# "Three Essays on the Measurement and Modeling of Subjective Well-Being"

**Dissertation**

**for the Faculty of Economics, Business Administration and Information Technology of the University of Zurich**

to achieve the title of
Doctor of Philosophy
in Economics

presented by

Raphael A. Studer
from Wangen bei Olten

approved in April 2013 at the request of
Prof. Dr. Rainer Winkelmann
Prof. Dr. Josef Zweimüller

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich hereby authorises the printing of this Doctoral Thesis, without thereby giving any opinion on the views contained therein.

Zurich, April 3, 2013

Chairman of the Doctoral Comittee: Prof. Dr. Dieter Pfaff

# Acknowledgements

During the years writing this thesis at the Chair for Statistics and Empirical Economic Research at the University of Zurich I have discovered Rainer Winkelmann's dedication to research and education. I have enormously benefitted from his creative ideas and profound advice and I cannot thank Rainer Winkelmann enough for his support and patience.

My grateful thanks are extended to Josef Zweimüller for many critical comments which improved my work and for co-advising this thesis. I am indebted to Steven Stillman, who introduced empirical economics to me. Arie Kapteyn had the kindness of inviting me to the RAND Corporation in Santa Monica CA for a research stay in fall 2012. I thank him and the staff of RAND Labor and Population for friendly welcoming me.

Financial funding from the Forschungskredit of the University of Zurich is gratefully acknowledged. I am also indebted to numerous collaborators of CentERdata at Tilburg University in the Netherlands, who implemented this thesis' ideas in the Longitudinal Internet Studies for the Social sciences.

With Gregori Baetschmann, Timo Boppart, Johannes Kunz, Kevin Staub and Andreas Steinhauer I had enthusiastic and critical fellows. Our discussions have led to fruitful collaboration and friendship; for both I am grateful.

Finally, I thank Lukas for sharing with me his indestructible ease of life and Andrea for her forthright criticism, infinite understanding and ongoing support.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

This dissertation contributes to the research into subjective well-being. It focuses on the measurement of happiness and the modeling of subjective well-being data in regression analyses. The dissertation consists of three chapters that provide new insights on these two topics by, first, implementing different answer scale designs for a happiness question in a randomized controlled experiment, second, recording response behaviors of respondents using computer-based survey technology and, third, developing new estimation strategies. Before each chapter's contributions to ongoing subjective well-being research are detailed, I briefly overview the process of how well-being surveys have become an integral part of empirical economic research.

## 1.1 Utility measurement - a short summary

Economics defines itself as the study of individuals' and societies' allocations of limited or scarce resources to satisfy needs and aspirations.[1] The prime example in microeconomics, for instance, is the optimization of a representative agent's consumption of goods constraint by its financial ability. To model the consumer's choices, the introduction of a utility function and a budget constraint is helpful. Let us focus on the former and assume a consumption set with two alternatives, good A and good B. A utility function, assigning a higher utility to good A than to good B, maps the economic agent's strict preference of good A over good B into numerical values. In this example the utility function is sufficient to carry ordinal information (Pareto, 1904). The optimal allocation of good A and good B is obtained by ranking the two alternatives only.

However, ordinal utility functions are limited in their use. Assigned utilities do not carry additional information on absolute utility levels. Utility levels, and therefore differences in utility levels of ordinal utility functions, are meaningless. Discounting utility levels over time periods and interpersonal comparisons are impossible. The concept of cardinal utility

[1] The definition dates back to Robbins (1935) and is generally found in introductory economic textbooks nowadays.

is needed.

The first to mention the possibility of a utility measurement is Bentham (1823), who proposed to assess a person's happiness by intensity and duration of emotional states such as pleasure or pain. Also Edgeworth (1881) imagined a perfect instrument constantly assessing the height of pleasure an individual experiences. However, such hedonometers remained thought experiments.

Further suggestions on the measurement of utility have been put into practice. Cardinal utility functions were derived from functional form assumptions based on economic axioms (Christensen et al. 1975). Well-being was approximated by monetary measures such as relative income (refer to Sen, 1979 for a review). Not only objective measures, but also subjectively assessed data have been used. The Leyden approach (refer to van Praag and Frijters, 1999 for a review), for instance, constructs a utility function through survey participants' perceptions of various income levels. Stated concisely, the development of utility measurements has concerned economists for a long time.

## 1.2 The emergence of happiness economics

Easterlin (1974) introduced answers to an explicit happiness question as a measure of well-being in empirical economic research. Such happiness questions are featured nowadays in the majority of household surveys and generally ask respondents to assess how happy or satisfied they consider themselves. Even though measuring an individual's utility through self-reported happiness seems to be a straightforward method, happiness remains a vague notion. Throughout the ages, numerous conceptualizations of happiness have been proposed (refer to Kesebir and Diener, 2008 for a review). A suitable definition for the purpose of this dissertation is given by Veenhoven (2009), who defines happiness as "the degree to which one evaluates one's life-as-a-whole positively" (p. 45). Regardless of the unwieldy concept of happiness, Easterlin's (1974) publication has drawn economists' attention to

self-reported well-being indicators.

In his seminal study, Easterlin (1974) concluded that the growth in per-capita income does not translate into higher average self-reported happiness across time or countries. Easterlin's paradoxical finding has been replicated and different explanations have been proposed.[2] Di Tella and MacCulloch (2006) review the hypotheses of happiness adapting to economic growth and of happiness reacting to relative rather than to absolute income. Further studies focus on individuals' incomes rather than aggregated income measures and identify a positive and statistically significant relationship between income and subjective well-being in either cross-section or panel data and with or without exogenous variation in income (Clark et al., 2008 and Kahneman and Deaton, 2010). These findings are in line with the assumption common in economic theory that higher income results in higher utility.

Happiness scores seem to provide some valid information on utility. However, the utility which happiness scores aim to measure is not of the same type as the utility introduced earlier. The utility function modeling the agent's optimal allocation of income to consumption of goods was employed to take an optimal decision reflecting the agent's preferences. Self-assessed happiness, though, evaluates an individual's life which results from various decisions. By measuring such "experienced" utility, happiness questions provide interesting data to estimate impacts of policies, economic shocks or life events on well-being.[3]

Research into subjective well-being has burgeoned in recent years. Studies have analyzed housing market equilibria (Stutzer and Frey, 2008) or calibrated economic parameters (Layard et al., 2008) based on happiness data. Deaton (2012) studies how the recent financial crisis affected life satisfaction of Americans. The effects of life events such as marriage, divorce or fertility decisions on subjective well-being have been studied using panel analyses (Clark et al., 2008b). Persons unemployed are found to report lower life satisfaction

[2] For a recent study that opposes the absence of an effect in cross-country analyses refer to Stevenson and Wolfers (2008).

[3] Bentham (1823) used the words utility and happiness already interchangeably. Only with Pareto (1904) utility was reduced to preference representations.

than employed individuals (e.g., Winkelmann and Winkelmann, 1998). This selective list of studies is by far not complete.

The popularity of self-reported well-being measurements in empirical economic research has stimulated methodological research as well. During recent years, regarded economic journals have published research on various methodological topics that arise with subjective well-being data. For example, non-linear panel estimators have been proposed for discrete happiness data (Ferrer-i-Carbonell and Frijters, 2004 and Baetschmann et al., 2011). Among others, Oswald and Wu (2010) studied the validity of single-item life satisfaction data. Kahneman and colleagues developed, in the spirit of Bentham (1823), a multiple item measurement of subjective well-being; the Daily Reconstruction Method (Kahneman et al., 2004). Doubtless, this research has contributed to develop happiness economics further.

Probably, the next level for subjective well-being data in empirical economics is the establishment of national accounts, as proposed by several researchers (Diener, 2000 and Kahneman et al., 2004). Some European countries are taking the first steps towards national subjective well-being accounts already. In the year 2008, the French government installed the Commission on the Measurement of Economic Performance and Social Progress chaired by Joseph Stiglitz, which concluded on the necessity of the development of a plural measurement of well-being combining objective and subjective indicators of quality of life (Stiglitz et al., 2009). Switzerland follows these recommendations. The recently developed indicator system of the Swiss Federal Statistical Office complements Gross Domestic Product with measures of monetary inequality and subjective well-being data.

Even though a huge interest into subjective well-being data is observable on all fronts, self-reported data is still facing skepticism. Manski (2000) summarizes some economists' attitudes towards self-reports by criticizing that his colleagues prefer to believe in individuals' actions rather than in individuals' words. In fact, economists' hostility to self-reports reflects *a priori* doubts about the reliability of subjectively assessed data (Bertrand and Mullainathan, 2001) and is rarely based on empirical evidence. The existence of es-

says assessing and improving the meaningfulness of self-reports is mostly being ignored by economists, whether they are in favor or reluctant to self-reported measures (Edwards, 2012).

Recent studies reveal that subjective well-being data are not beyond reproach. Vignette studies show that response scales are subject to idiosyncratic rescaling (Kapteyn et al., 2010). Randomized controlled experiments suggest self-reports of subpopulations to be affected by the labeling of scales (Conti and Pudney, 2011) and question orders to be susceptible to impact responses (Deaton, 2012). Moreover, panel learning effects are found to systematically reduce levels of reported happiness of an individual over time (van Landeghem, 2012). On one hand this research provides important insights on subjective well-being measures and possible shortcomings in their use. On the other hand these critical findings highlight the need to develop practical methods which are able to uncover true well-being or utility from self-assessed measures.

## 1.3 Dissertation overview

This dissertation contributes to the ongoing research that promotes the understanding and accurate use of subjective well-being data. The three chapters are independent studies. Their common goal is to identify and solve methodological challenges that arise with happiness data. The novel insights presented in this dissertation demonstrate how commonly accepted findings of subjective well-being studies cast on measurement methods, response processes and estimation strategies.

Chapters 2 and 3 employ computer-based survey technology in order to study the measurement of happiness. Chapter 2 researches in a randomized controlled experiment a new continuous measurement of happiness that proves to be a valid alternative to the generally used discrete rating scale. Chapter 3 records reporting behaviors, which for example measure the degree of meaningfulness of happiness answers, and tests happiness reporting

functions in order to detect how response behaviors affect reported happiness levels and perceptions of happiness determinants. Both chapters make use of a non-linear regression model to estimate happiness equations. For this reason, the Rating Scale Model is introduced hereafter as well. In contrast to the foregoing two chapters, Chapter 4 is not primarily concerned about the measurement. The application to the effect of motherhood on life satisfaction illustrates that the development and use of accurate empirical models for subjective well-being data is essential. The regression analyses proposed in this chapter account for censoring and self-selection of mothers and resulting estimates contradict earlier evidence.

### 1.3.1 Summary Chapter 2

Answers to happiness questions have been interpreted as providing valid information on the utility of individuals. Studies doing so employ happiness data gathered by a discrete single-item scale, ranging from 0 to 10 for instance. However, the discrete scale drives a wedge between the underlying latent dimension and the measurement thereof. In fact, some empiricists have opposed discrete happiness scales to be of ratio quality and interpreted the scale as ordinal. This implies that differences between answer categories are not defined anymore. But if discrete happiness scales provide information about a ranking only, the subtext of happiness measurements is swept off. In order to estimate the impact of economic context on overall well-being or to quantify intertemporal and interpersonal differences in utility levels, ordinal data is not sufficient. A continuous measurement scale overcomes this statistical issue. Chapter 2 introduces the visual analogue scale in a representative online survey.

The visual analogue scale is a bounded line. Respondents report levels of happiness by choosing a point on the line. It is argued that the line attributes distances between individuals' choices a meaningful visual interpretation. Using computer-based surveys, distances between answers can be covertly measured on a continuous scale. If one is willing

to interpret subjective well-being data as providing information on individuals' utility, the continuous scale leads to a truly cardinal utility measure.

In order to compare the innovative continuous measurement to the established discrete measurement, both happiness scales were implemented in a randomized controlled experiment in the Dutch Longitudinal Internet Studies for the Social Sciences. The experiment generates two groups of respondents with the same latent distributions of true happiness such that differences in distributions of scores or strengths of happiness determinants are fully attributable to scale design.

Results suggest that respondents are more likely to score closer to the extremes on the continuous scale. Implications are twofold. First, if the visual analogue scale is interpreted as the true replication of latent happiness, the finding suggests that respondents perceive the interval lengths between discrete response categories as unequal. In fact, the thought experiment indicates that on the discrete scale intervals between extreme categories, and especially between the lower categories, are perceived narrower than the ones between middle categories. Second, related to the first implication, high frequency categories present with the use of discrete scales, usually the scores of 7 and 8, are artifacts of too little discriminating power. Thus, the visual analogue scale leads individuals to overcome the endpoint aversion likely to be present with a discrete measurement and move closer to the extremes. Additionally, the experiment reveals an increased likelihood for women of revising scores downwards on the visual analogue scale making the gender inequality, apparent in discrete subjective well-being data, to disappear.

The conclusions of Chapter 2 are threefold. First, subjective well-being is considered to be a continuous phenomenon and therefore the discrete scale is a problematic measurement as it leads to a discretization of the underlying dimension and foils economists' intentions to measure cardinal utility. The continuous visual analogue scale is a valid substitute that overcomes this concern. Second, the visual analogue scale's scores exhibit higher variances, thus provide more information on subjective well-being differences. Third, a difference in

well-being among genders strongly depends on the scale employed. The implication of this result is unclear, as the true underlying dimension is not measurable.

### 1.3.2 Summary Chapter 3

Despite its popularity, subjective well-being measurements are still facing skepticism. A main reason for this objection may be the reporting black-box. Opposed to researchers using objective measurements the trustworthiness of self-reported data is unknown. Do survey participants report on the happiness question seriously or are happiness scores based on thoughtless answers? Chapter 3, which is joint work with Rainer Winkelmann, enlightens the reporting black-box.

To learn more about the reporting behavior of participants, surveys can ask respondents to self-assess cognitive effort invested into a happiness question. However, with the development of computer-based surveys novel technology is available to objectively record response behaviors of participants. These response behaviors provide information about an answer's meaningfulness. Time stamp data, for example, enable the recording of response times for the happiness question as well as the recording of the day or time participants responded to different question modules. Information about three reporting behaviors (answer speed, self-reported cognitive effort and questionnaire order) were collected in the Dutch Longitudinal Internet Studies for the Social Sciences. Chapter 3 researches whether these reporting behaviors affect levels of reported happiness and whether the sensitivity of reported happiness to socioeconomic characteristics varies with these response behaviors. The estimation results are employed to test the existence of different happiness reporting functions and provide insights to which degree reported happiness is useful to uncover the latent happiness.

Results suggest that slower responses and higher self-stated cognitive effort are associated with lower reported happiness. In multivariate happiness equations, these factors moderate the estimated effect of income on happiness as well, while no interaction effects

are found for the remaining socioeconomic determinants of happiness. Moreover, an explanation is provided how response times relate to momentary mood of survey participants. Cognitive psychology suggests that slower respondents are more likely to report lower happiness than a comparable person, because of worse mood.

The conclusions presented in Chapter 3 are twofold. First, reporting behaviors are easily measurable with modern survey technology and enrich happiness equations by explaining residual variance. Second, the measurement of subjective well-being as well as the estimation of associations between happiness and material determinants depend on reporting correlates. In light of these results, our model suggests that estimated trade-off ratios reported in happiness studies are not able to map well-being or utility trade-off ratios.

### 1.3.3 The Rating Scale Model employed in Chapters 2 and 3

Chapters 2 and 3 both employ a novel regression model that is shortly introduced here. The Rating Scale Model, of which an early version was published by Studer and Winkelmann (2011), is in the spirit of Papke and Wooldridge (1996) who proposed a model for fractional dependent variables. However, such a model has not yet been applied to a rating scale variable like self-assessed happiness.

The wide use of discrete happiness scales has led to the development of ordered regression models. However, many happiness economists continue to interpret subjective well-being data as being of ratio level and study the effects of happiness determinants on subjective well-being by means of linear regression. This empirical strategies enables a simple computation of the average treatment effect. Moreover, given the state of the art, the linear regression framework would be considered as the appropriate empirical model for a continuous happiness measurement like the one introduced in Chapter 2.

However, the linear regression model has one major drawback; if it is employed to model a rating as dependent variable. Rating variables such as subjective well-being are limited dependent variables. In other words, ratings are defined on a subset of the real line,

namely between the best and the worst rating. The linear right hand side of the regression model, however, is defined over the entire real line and does not respect the bounds of the rating variable. Ordinary least squares results in improbable constant marginal effects and predictions outside the logically possible range.

The Rating Scale Model's goals are twofold. First, it shall overcome the linear regression's model inconsistency. Second, the new model shall provide empiricists a simple method to compute average treatment effects. For the sake of the latter, the Rating Scale Model specifies a conditional expectation based on a single index of explanatory variables. In order to account for the boundedness of the dependent rating variable the conditional expectation function is chosen to be non-linear. Only a non-linear function is able to map the linear index of explanatory variables on the rating variable's support.

The Rating Scale Model employed in Chapters 2 and 3 of this dissertation models subjective well-being score $y_i$ for independent survey participants $i = 1, \ldots, N$ with domain $y \in [0, y^{max}]$. The Rating Scale Model is defined by a non-linear conditional expectation function:

$$\mathrm{E}(y_i|x_i) = G(x_i'\beta), \tag{1.1}$$

such that $0 \leq G \leq y^{max}$. $G$ is a monotonic and twice differentiable function. The linear index $x_i'\beta$ consists of a vector $x_i$ and a parameter vector $\beta$ both of dimension $(k \times 1)$.

$G$ can be specified parametrically or estimated semiparametrically. In Chapters 2 and 3 of this dissertation the cumulative density function is assumed to be the logistic distribution multiplied by $y^{max}$:

$$G(x_i'\beta) = y^{max}\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}. \tag{1.2}$$

One method to estimate the vector $\beta$ in such a parametric Rating Scale Model is quasi-maximum likelihood estimation. The conditional expectation function (1.1) has to be embedded in a suitable distribution of the linear exponential family (Gourieroux et al., 1984). If the Bernoulli distribution $B(1, p)$ is used for estimation, one needs to observe

that $0 \leq p \leq 1$. This condition is satisfied once both sides of equation (1.1) are divided by $y^{max}$. The Bernoulli quasi-maximum likelihood estimator solves:

$$\sum_{i=1}^{N} \frac{(y_i - G(x_i'\beta))}{y^{max}} \frac{g(x_i'\beta)}{(1 - G(x_i'\beta)/y^{max})G(x_i'\beta)} x_i = 0 \text{ where } g(x_i'\beta) = \frac{\partial G(x_i'\beta)}{\partial x_i'\beta}. \quad (1.3)$$

The asymptotic distribution of the quasi-maximum likelihood estimator is normal with variance-covariance matrix in the usual sandwich form:

$$\text{AVar}(\hat{\beta}) = N^{-1} I^{-1}(\beta) J(\beta) I^{-1}(\beta), \quad (1.4)$$

where

$$I(\beta) = -\text{E}[H(\beta; y, x)] = -\text{E}\left[\frac{-g(x_i'\beta)^2/y^{max}}{(1 - G(x_i'\beta)/y^{max})G(x_i'\beta)} x_i x_i'\right] \quad (1.5)$$

and

$$J(\beta) = \text{Var}(s(\beta; y, x)) = \text{E}\left[\left(\frac{y_i - G(x_i'\beta)}{y^{max}}\right)^2 \frac{g(x_i'\beta)^2}{(1 - G(x_i'\beta)/y^{max})^2 G(x_i'\beta)^2} x_i x_i'\right]. \quad (1.6)$$

The Rating Scale Model's marginal effects are not constant. The average marginal effect for the $l$-th regressor $x_l$ is given by $\beta_l \, \bar{y}(y^{max} - \bar{y})/y^{max}$, where $\bar{y}$ is the average predicted conditional expectation function.

In order to illustrate the implications of the linear regression's model inconsistency suppose that a rating variable $y_i$ has support [0,10] and is equal to:

$$y_i = 10 \cdot \frac{exp(x_i'\beta)}{1 + exp(x_i'\beta)} + u_i \, \text{for i} = 1, \ldots, 10'000. \quad (1.7)$$

The disturbance term $u_i$ is drawn from a Normal distribution with expectation 0 and variance: $\frac{exp(x_i'\beta)}{1+exp(x_i'\beta)} \left(1 - \frac{exp(x_i'\beta)}{1+exp(x_i'\beta)}\right)$. The linear index is of the form:

$$x_i'\beta = \beta_0 + \beta_1 \cdot x_{i1} = 0.5 + 2 \cdot x_{i1}, \text{with x}_{\text{i1}} \sim \text{Uniform}[-8, 8]. \quad (1.8)$$

Table 1.1 reports estimates obtained by ordinary least squares and Bernoulli quasi-maximum likelihood estimation. The parameters $\beta_0$ and $\beta_1$ and the average effect of $x_1$ are displayed. The first column presents the true parameters of the data generating process. Parameter estimates in the second column for the linear regression model are found to be

biased. Ordinary least squares neither estimates the parameters nor the average effect of the data generating process. Estimates presented in the last column reveal that the Rating Scale Model results in unbiased point estimates and consequently in an unbiased average marginal effect of $x_1$ on $y$.

Figure 1.1 shows both regressions' mean predictions for the rating dependent variable. In the left graph mean predictions hurt the bounds of 0 and 10. Such impossible predictions happen for one third of all observations in the linear model. In the right graph, the Rating Scale Model's predictions are plotted. The dependent rating variable's support is not violated.

### 1.3.4 Summary Chapter 4

Chapter 4, joint work with Gregori Baetschmann and Kevin E. Staub, differs in its focus from the previous two chapters. It presents an empirical application of discrete life satisfaction data rather than research into the measurement thereof and does so by employing panel data rather than cross-section data. However, Chapter 4 is also concerned about the use of accurate empirical strategies to model subjective well-being data.

Happiness economists frequently employ panel data. On one hand, panel regressions can control for time invariant unobserved characteristics such as personality traits which are found to be an important determinant of subjective well-being (Ferrer-i-Carbonell and Frijters, 2004). On the other hand, panel data is particularly interesting to study how life events affect an individual's well-being. Chapter 4 focuses on women's fertility decisions and shows that standard estimation techniques are not sufficient to obtain unbiased estimates of the effect of motherhood on life satisfaction.

Previous studies have estimated satisfaction differences between parents and comparable childless adults, mostly finding small and often negative effects of parenthood. This stands in sharp contrast to standard theoretical economic models of fertility which assume that the net utility gain of motherhood is positive. However, Chapter 4 shows that such empirical

comparisons of ex-post similar individuals are problematic for various reasons. The most important is the existence of self-selection into motherhood. For instance, 70% of women who gave birth to a child during the years 2002 to 2009 reported in the German Socio-Economic Panel that their first birth was planned. In fact, non-mothers' and mothers' satisfaction paths diverge around five years before mothers' first birth, even after adjusting for differences in socioeconomic characteristics. Chapter 4 examines the selection issue in detail by exploiting the extended longitudinal dimension of the German Socio-Economic Panel to track self-reported life satisfaction of women eventually to become mothers and of women eventually attaining a completed fertility of zero.

Three empirical strategies are proposed to account for selection. A nearest neighbor matching estimator pairs the mothers to the most similar non-mothers in terms of pre-birth covariates and pre-birth life satisfaction. A regression controls for pre-birth covariates and pre-birth life satisfaction trend and level. The third approach exploits intrapersonal variation only. A fixed effect regression with dummy variables for the last five pre-birth years is estimated. Does the use of these empirical strategies change the sign and size of the effect of motherhood on life satisfaction?

All three regressions result in a similar estimate of the effect of motherhood on life satisfaction. A long lasting positive effect is found. The maximum life satisfaction difference between mothers and non-mothers is reached in the year of delivery and the difference remains positive during the first twenty years of the child's age. Moreover, evidence reveals heterogeneous effects for different ages at first birth and for different numbers of children.

The conclusions presented in Chapter 4 suggest that cross-section and panel studies analyzing the effect of life events on individuals' life satisfaction paths are susceptible to self-selection of individuals into such events. By studying the effect of motherhood on life satisfaction, Chapter 4 demonstrates how important the development of suitable empirical strategies is. The negative effect of motherhood on life satisfaction reported in earlier studies turns out to be positive, if appropriate data and estimation specifications, which

account for the censoring of mothers and the selection into motherhood, are employed. This result is in line with a neo-classical view of choice behavior based on utility maximization.

# References


Baetschmann G., K. E. Staub and R. Winkelmann, 2011, "Consistent Estimation of the Fixed Effects Ordered Logit Model", *IZA Discussion Paper*, No. 5443

Bentham, J., 1823, *An introduction to the principles of morals and legislation*, London: W. Pickering and R. Wilson

Bertrand, M. and S. Mullainathan, 2001, "Do People Mean What They Say? Implications for Subjective Survey Data", *The American Economic Review*, 91, 2, 67-72

Christensen, L R., D. W. Jorgenson and L. J. Lau, 1975, "Transcendental logarithmic utility functions", *American Economic Review*, 65, 367-383

Clark, A. E., P. Frijters and M.A. Shields, 2008, "Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles", *Journal of Economic Literature*, 46, 1, 95-144

Clark A. E., E. Diener, Y. Georgellis and R. E. Lucas, 2008b, "Lags and leads in life satisfaction: a test of the baseline hypothesis," *Economic Journal*, 118, 529, F222-F243.

Conti, G. and S. Pudney, 2011, "Survey Design and the Analysis of Satisfaction", *The Review of Economics and Statistics*, 93, 3, 1087-1093

Deaton, A., 2012, "The Financial Crisis and the Well-Being of Americans", *Oxford Economic Papers*, 64, 1-26

Diener, E., 2000, "Subjective Well-Being: The Science of Happiness and a Proposal for a National Index", *American Psychologist*, 55, 1, 34-43

di Tella, R. and R. MacCulloch, 2006, "Some Uses of Happiness Data in Economics", *Journal of Economic Perspectives*, 20, 1, 25-46

Easterlin, R. A., 1974, "Does Economic Growth Improve the Human Lot?", In P. A. David and M. W. Reder (Eds.), *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, New York: Academic Press, Inc.

Edgeworth, F., 1881 [1961], *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*, New York: A.M. Kelly.

Edwards, J. M., 2012, "The history of the use of self-reports and the methodology of economics", *Journal of Economic Methodology*, 19, 4, 357-374

Ferrer-i-Carbonell, A. and P. Frijters, 2004, "How important is methodology for the estimates of the determinants of happiness?", *Economic Journal*, 114, 641-659

Gourieroux, C., A. Monfort and A. Trognon, 1984, "Pseudo Maximum Likelihood Methods: Theory", *Econometrica*, 52, 3, 681-700

Kahneman, D., A. B. Krueger, D. Schkade, N. Schwarz and A. Stone, 2004, "Toward National Well-Being Accounts", *The American Economic Review Papers and Proceedings*, 94, 2, 429-434

Kahneman, D. and A. Deaton, 2010, "High income improves evaluation of life but not emotional well-being", *PNAS Early Edition*, published online September 6 2010, DOI 10.1073/pnas.1011492107

Kapteyn, A., J. P. Smith and A. van Soest, 2010, "Life Satisfaction", In E. Diener, D. Kahneman and J. Helliwell (Eds.), *International Differences in Well-Being*, Oxford University Press, New York

Kesebir, P. and E. Diener, 2008, "In Pursuit of Happiness: Empirical Answers to Philosophical Questions", *Perspectives on Psychological Science*, 3, 2, 117-125

Layard, R., G. Mayraz and S. Nickell, 2008, "The marginal utility of income", *Journal of Public Economics*, 92, 1846-1857

Manski, C. F., 2000, "Economic Analysis of Social Interactions", *The Journal of Economic Perspectives*, 14, 3, 115-136

Oswald, A. and S. Wu, 2010, "Objective Confirmation of Subjective Measures of Human Well-Being: Evidence from the U.S.A.", *Science*, 327, 576-579

Papke, L. E. and J. M. Wooldridge, 1996, "Econometric methods for fractional response variables with an application to 401(k) plan participation rates", *Journal of Applied Econometrics*, 11, 6, 619-632

Pareto, V., 1904, *Manuel d'conomie politique*, *Social Behavior and Personality*, Paris: M. Girard

Robbins, L., 1935, *An Essay on the Nature and Significance of Economic Science*, London: Macmillan

Sen, A., 1979, "The Welfare Basis of Real Income Comparisons: A Survey", *Journal of Economic Literature*, XVII, 1-45

Stevenson, B. and J. Wolfers, 2008, "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox", *Brookings Papers on Economic Activity*, Spring

Stiglitz, J., A. Sen and J-P. Fitoussi, 2009, "Report by the Commission on the Measurement of Economic Performance and Social Progress", http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf, last consultation 21.11.2012

Studer, R. and R. Winkelmann, 2011, "Specification and Estimation of Rating Scale Models – with an Application to the Determinants of Life Satisfaction", *Working paper series, Department of Economics University of Zurich*, No. 3

Stutzer, A. and B. S. Frey, 2008, "Stress that Doesn't Pay: The Commuting Paradox", *Scandinavian Journal of Economics*, 110, 2, 339-366

van Landeghem, B., 2012, "Panel Conditioning and Self-reported Satisfaction: Evidence from International Panel Data and Repeated Cross-sections", *SOEPpaper*, No. 484

van Praag, M. S. and P. Frijters, 1999, "The Measurement of Welfare and Well-Being: The Leyden Approach", In D. Kahneman, E. Diener and N. Schwarz (Eds.), *Well-Being The Foundations of Hedonic Psychology*, New York: Russel Sage Foundation

Veenhoven, R., 2009, "How do we assess how happy we are? Tenets, implications and tenability of three theories", In A. K. Dutt and B. Radcliff (Eds.), *Happiness, Economics and Politics: Towards a multi-disciplinary approach*, Cheltenham: Edward Elger Publishers ISBN 978 1 84844 093 7, Chapter 3, page 45-69

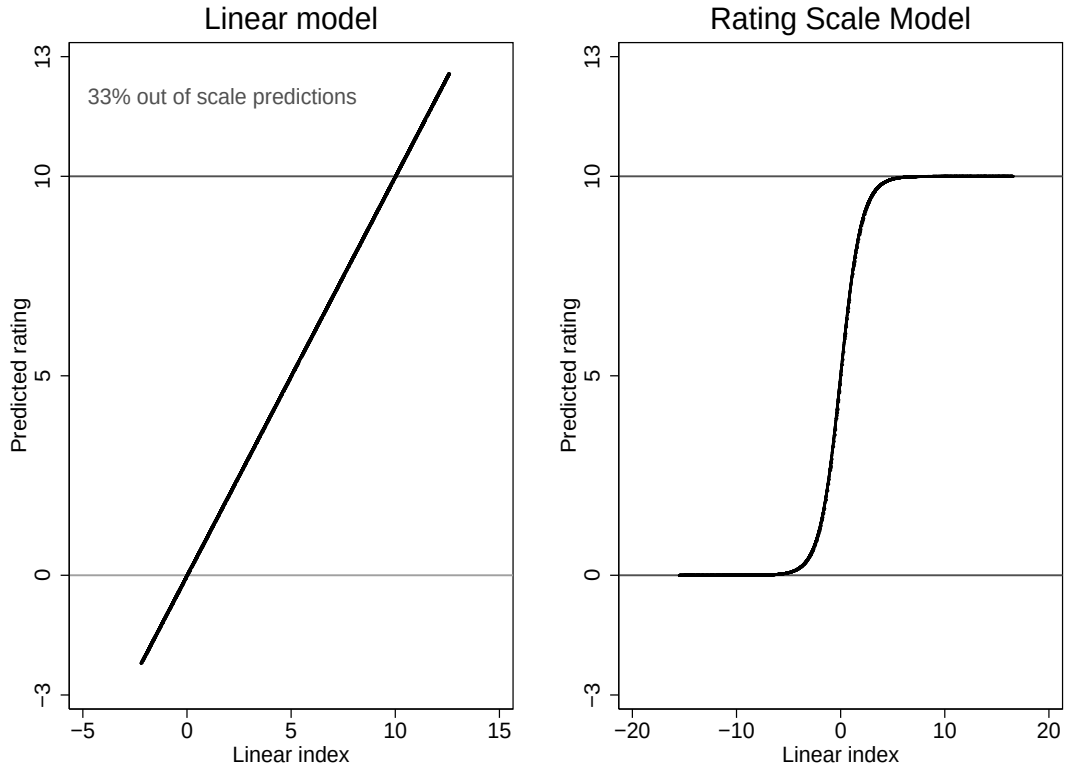
Winkelmann, L. and R. Winkelmann, 1998, "Why Are the Unemployed So Unhappy? Evidence from Panel Data", *Economica*, 65, 257, 1-15

Table 1.1: Rating Scale Model illustration - Point estimates

| | DGP | Linear regression | Rating Scale Model |
|---|---|---|---|
| $\beta_0$ | 0.5 | 5.190 | 0.504 |
| | | (0.019) | (0.002) |
| $\beta_1$ | 2 | 0.924 | 2.000 |
| | | (0.004) | (0.002) |
| Average marginal effect for $x_1$ | 0.616 | 0.924 | 0.616 |
| | | (0.004) | (0.001) |

*Notes:* Robust standard errors in parentheses. The Rating Scale Model uses Bernoulli quasi-maximum likelihood estimation with a logit type link function.

Figure 1.1: Rating Scale Model illustration - Predicted mean rating



# Chapter 2

# Does it matter how happiness is measured? Evidence from a randomized controlled experiment

Similar versions of this chapter have been published in the *Working Paper Series* of the Department of Economics, University of Zurich, No. 49 and in the Journal of Economic and Social Measurement, **37**(4), pp. 317-336.

*Acknowledgements*: I thank Rainer Winkelmann for extensive discussion and advise and Joshua Angrist, Luke Connelly, John Haisken-DeNew, Steven Stillman and Joseph Zweimüller for important comments. Many useful remarks from seminar participants at the Universities of Geneva and Zurich, the 2012 Workshop in Labor Economics and the 2012 conferences of the Royal Economic Society and European Society for Population Economics are gratefully acknowledged. This paper draws on data of the LISS panel of CentERdata.

## 2.1 Introduction

Interest in the determinants of subjective well-being using survey data has burgeoned in recent years. Research contributes to the questions whether socioeconomic factors, such as income or unemployment, and sociodemographic factors, like marriage or parenthood, affect well-being (see for reviews Kahneman et al., 1999 and Frey and Stutzer, 2002). Presented evidence is generally based on happiness or life satisfaction data self-assessed by survey participants on a Likert scale (LS) (Likert, 1932). The discrete rating scale is widely accepted and little has been done to find alternative measurements of subjective well-being. This study proposes a continuous rating scale to measure individual happiness, the visual analogue scale (VAS).

The LS and VAS were implemented in a Dutch representative online survey. Respondents answered both scales with a one month washout time between the two assessments. The order of scales was randomized in order to generate two groups of respondents with equal latent well-being distributions. Comparison of happiness answers among these two groups provides novel insights on how scale design affects inferences drawn from subjectively measured well-being data. The results suggest stylized findings in happiness research to be particular to the discrete happiness scale. Moreover, while associations between happiness and socioeconomic characteristics are robust among scales, subpopulations defined by gender are found to react differently to the question design. In turn this affects group comparisons of happiness scores between women and men. These results challenge the habit of employing discrete happiness scales.

The discrete single-item happiness question, nowadays featured by virtually every well-being survey, was proposed by Fordyce (1987). Its qualities have been widely studied. The effects of labels (Larsen et al., 1984) and numbers (Cummins, 2003) of categories on answers have been examined, for instance. But the accurateness of other rating scales has been tested in subjective well-being research (Diener, 1994) as well. Andrews and Crandall (1976), for instance, assessed data quality of faces and ladder scales. However, these rating

scales remained discrete. With new computer-based survey technology novel measurements that overcome shortcomings of the discrete scale become available.

Van Praag and Ferrer-i-Carbonell (2004) note that individuals likely perceive satisfaction as a continuous phenomenon bounded by the states of complete dissatisfaction and complete satisfaction. That happiness is continuous is implicitly accepted in every empirical application estimating a happiness regression. Ordinary least squares specifies the conditional mean of happiness scores continuously and ordered response models build on a continuous latent framework. The LS therefore implies a discretization of the underlying true happiness into a discrete score. On one hand, this may lead to systematic transformation error. On the other hand, the resulting discrete scores may carry ordinal information only (Kristoffersen, 2010). The latter concern is particularly harmful in economics where subjective well-being data are sometimes interpreted as a cardinal measurement for utility. The VAS overcomes these shortcomings. It acts as a reference continuum for the latent happiness and attributes distances between individuals' choices a meaningful visual interpretation.

The VAS (Hayes and Patterson, 1921) is simply a bounded line. Respondents assess their happiness by setting a marker on the VAS. The VAS has been extensively used in medical pain research (McCormack et al., 1988) and tested against discrete scales (e.g., Lara-Muñoz et al., 2004). A recent literature has compared LS and VAS in various computer based experiments (Couper et al., 2006). Evidently, only four happiness studies have used the VAS (Matsubayashi et al., 1992; Saris et al., 1998; Bouazzaoui and Mullet, 2002; Hofmans and Theuns, 2008). Saris et al. (1998) did not declare how the "graphical line scale" was implemented and for which satisfaction domain it was used. The other three articles are small sample paper and pencil vignette studies, which do not have any counterfactual for the VAS scores, i.e. LS scores for the same individuals. No evidence exists how a discrete and a continuous scale differ in assessing subjective well-being.

The chapter starts by presenting the survey and question design and assessing the

quality of the experiment. Section 2.3 reviews the existing research on comparison of single-item happiness scales and provides estimates for reliability and validity measures. The analyses suggest that happiness data produced by either measure are valid and reliable. Distributional analyses are presented in Section 2.4. The experiment shows lower average and wider spread happiness scores for the VAS. This finding is not caused by a rescaling of the measurements. In fact, higher order standardized moments suggest an increased likelihood of VAS scores closer to the scale's end-points. The unexplained pattern of LS high frequency categories may simply be due to too little discriminating power. Section 2.5 exploits the existence of two parallel happiness questions to investigate the impact of rating scales on correlates of happiness for a common set of respondents. The significant gender gap which is present when LS data are used vanishes with the use of VAS data. Especially, female respondents are found to change happiness reports following which scale was used. Section 2.6 presents the conclusions.

## 2.2 Survey design

The randomized controlled experiment that is used in this paper was implemented in the Longitudinal Internet Studies for the Social Sciences (LISS) panel. The LISS panel was established by CentERdata based at Tilburg University in the Netherlands. 10'150 random addresses were drawn from a 10% sample of the Dutch population register. The oldest inhabitant at each address was approached via mail, with a letter and incentive payment of 10 Euros. In case of non-response, the person was called or visited. 5176 households agreed to participate in the survey. Households without a broadband internet connection or computer were provided with it. During the first survey year in 2007, the average monthly answer rate was 73% of all members of participating households (Scherpenzeel, 2009). Knoef and de Vos (2009) concluded that the initial sample contained an underrepresentation of elderly people and of some ethnicities. In 2009, a refreshment sample stratified by age,

ethnicity and household types was successful in establishing representativeness of the LISS panel (de Vos, 2010).

An e-mail at the beginning of a month invites participants to respond to the LISS panel. The response burden of each wave differs. Survey respondents can choose on which day and at what time they want to answer each of the questionnaires. Three types of questionnaires can be distinguished. The Background Variable questionnaire is sent every month to the contact person of the household, but needs only to be updated if any changes in the core socioeconomic or sociodemographic variables, such as income, education, age, civil status or household composition, occurred for any household member. Ten Core Studies, for instance on health or religion, are repeated once a year and sent in two subsequent months in order to maximize response. Core Studies are not equally distributed over the year. During a specific month none or several Core Studies may be included in the survey. Assembled Studies, like the experiment analyzed in this paper, are one-off studies and sent to all participants as well.

The experiment was implemented during the survey months March and April 2011. The web link to the Assembled Study directed participants to a single-item happiness question. Answers had to be given either on a LS or a VAS. Answer scales were randomly assigned in March at the moment people opened the questionnaire. In the subsequent month, the scales were changed or again randomly assigned if people had not answered during the March wave. In the best case, every survey participant reported his or her happiness using the VAS and the LS. This crossover design has two advantages. First, the dependent sample increases power of test statistics. Second, any time effects that occur in subgroups of the sample are captured in both scales equally and do not distort the analysis.

The crossover experiment is summarized in Figure 2.1. In the first wave 5042 individuals and in the second wave 4795 individuals participated. 4274 subjects responded in both waves. 1681 observations are lost due to missing information on background variables. Furthermore, 6 individuals had to be dropped from the data set as they opened the

questionnaire twice during one month. In May 2011, the month after the experiment took place, the LISS Core Study was dedicated to a personality questionnaire. The LISS panel Personality Study gathers not only information on overall happiness on a LS ranging from 0 to 10 (11 points) but also on personality traits, like emotional stability or self- esteem. 5230 individuals responded to the Personality Study in May, out of which 3770 individuals had already assessed their happiness in March and April. Data of the March and April waves will be uniquely used to quantify differences in distributions of scores (Section 2.4) and happiness correlates (Section 2.5). For the assessment of data quality (Section 2.3) data of the May wave will also be employed.

Screenshots of the two questions implemented in the experiment are presented in Figure 2.2. The LS ranges from 0 to 9 (10 points). This question design is used in the World Values Survey the European Social Survey and the European Quality of Life Survey, for instance. The VAS is a continuous line. It neither carries numbers nor does it show categories. The reason for not showing markers on the VAS was to avoid focal points in order to maximize the variation in answers and to eliminate explicit display of intervals. Hence, possible differences in scores between the two measures would occur through changes in two languages of communication: graphical and numerical (Dilman, 2007). In order to prevent a possible systematic rescaling of the VAS the same endpoint labels as on the LS were used. For practical purposes a scale unit has to be chosen for the VAS. In this application VAS scores were covertly measured from 0 to 99. Therefore, the VAS measurement has ten times more discriminating power than the LS measurement.

Figure 2.2 shows that the implementation of both scales was otherwise identical: No questions were asked before the happiness question; the length of both scales was approximately equivalent; the VAS had no default marker to avoid artificial high frequency regions (Treiblmaier and Filzmoser, 2011); both scales were aligned horizontally, however, results should not differ to vertical scales (Funke et al., 2010; Paul-Dauphin et al., 1999); and the same anchor words were used for the LS and the VAS in order to avoid wording effects

(Weng, 2004). No hidden factors should cause any differences in response behavior.

In order to examine the question design, participants answered 5 evaluation questions after participating in the experiment. Difficulty in answering, clearness of the question, degree of thought provocation, interest and joyfulness were rated on a LS ranging from 1 (certainly not) to 5 (certainly yes). Figure 2.3 gives the distributions by scale types for all five evaluation questions. Distributions are very similar. A Pearson's Chi-squared test cannot reject the hypothesis of equality of distributions for the variables difficulty and joy. Even though for the remaining three evaluation questions equality of distributions is rejected, densities in all five categories do not differ by more than 1 percentage point. Given this evidence, it is reasonable to conclude that the question design affects answers not through response difficulties caused by one or the other rating scale.

Two concerns about the experiment may still be raised. First, screen resolution may differ among survey participants. A lower resolution leads to a wider VAS or LS. Previous empirical findings, however, suggest no effect of varying length of the VAS (Kreindler et al., 2003). Second, people can decide on the order of the three questionnaires each month on their own. Contact persons could have answered first the Background Variable questionnaire and second the happiness experiment. Order of questions have been shown to have important effects on answers (Schumann and Presser, 1981). Therefore, time stamp data was collected enabling to test and control for questionnaire order.

Tables 2.1 and 2.2 examine the quality of the present experiment. Table 2.1 evaluates whether the subsamples are truly random. The means of ex-ante characteristics are compared by scale types. Equality of means for almost any of the variables cannot be rejected by a t-test. Only the first moments of the variables age, marital status, working and citizenship differ significantly. However, the means are very similar in magnitude for the two groups and differ by 1 year or 3 to 4 percentage points. In the April wave mean equalities for the variables age and employment status cannot be rejected anymore. If randomization seems not complete on statistical grounds, the well-balanced samples suggest that it is practically. Table 2.2 reports estimates of the parameters capturing a time or questionnaire

order effect, if existent, i.e. the estimates of the following model:

$$s_{ijt} = g_j\left(\beta_{0j} + \beta_{1j} \cdot april_{ijt} + \beta_{2j} \cdot experiment2^{nd}_{ijt}\right) + u_{ijt} \tag{2.1}$$

The dependent variable $s_{ijt}$ is the happiness score of individual $i$ using rating scale $j$ in wave $t$. Depending on the scale that was employed, $s_{ijt}$ ranges from 0 to 9 or 0 to 99. The function $g$ accounts for the boundedness of the dependent variable and is specified as: $g_j(.) = s_j^{max}\frac{exp(.)}{1+exp(.)}$. In case of the LS scores $g$ is bounded between 0 and 9 and in case of the VAS mean predictions cannot exceed 0 or 99. The parameter $\beta_1$ captures a potential time effect, which is allowed to differ by scales. The variable $experiment2^{nd}_{ijt}$ takes the value 1 if, during the 2 hours preceding the response of the happiness question, the Background Variable questionnaire was opened by the contact person. Also this questionnaire order effect may vary by scales. The non-linear model is consistently estimated by quasi-maximum likelihood employing the Bernoulli distribution. The estimation and properties of such a Rating Scale Model are detailed in Section 1.3.3 of this dissertation. Table 2.2 shows the estimated average discrete effects resulting from the derivation of model (2.1) with respect to either variable. No time effect for any of the two scales is found. While questionnaire order does not matter for LS scores, it reduces VAS scores significantly, even though the effect remains small. Later analyses will control for the order of questionnaires.

Summing up, randomization was successful and no time effect distorts distributions of happiness scores. Given this evidence, differences in distributions found between the randomly assigned groups are interpreted below as rating scale design effects.

## 2.3 Data quality, validity and reliability of happiness scales

A huge body of literature stemming from different scientific domains has been interested in the quality of data produced by rating scales. This study's unique large scale experimental

set-up is a novel contribution to this literature. In this section I review different methods, including the true score model, analyses of response behavior and validity and reliability measurements.

## Data quality

A simple quality measurement of rating data is provided by the true score model (e.g., Saris and Gallhofer, 2007). Consider the observed score $s_i$ for individual $i$ being a noisy measure of the transformed score $t_i$: $s_{ij} = t_{ij} + \zeta_{ij}$. If the transformation for every rating scale $j$ is a linear function of the latent happiness $h_i$: $t_{ij} = v_j \cdot h_i + \eta_{ij}$, then substitution yields: $s_{ij} = v_j \cdot h_i + \epsilon_{ij}$. The three parameters of interest $v_j^2$ for all three rating scales at hand ($j = \{vas, ls10, ls11\}$) are identified through the three correlations between the different $s_{ij}$'s. In fact $corr(s_{i,vas}; s_{i,ls10}) = \frac{v_{ls10} \cdot v_{vas} \cdot Var(h_i)}{\sqrt{Var(s_{i,ls10}) \cdot Var(s_{i,vas})}}$, reduces to $corr(s_{i,vas}; s_{i,ls10}) = v_{ls10} \cdot v_{vas}$ when each scale's variance is standardized to unity. The lowest quality is found for the VAS (0.67), followed by the 10 points LS (0.69) and the 11 points LS (0.71). However, the true score model is not entirely applicable, as it erroneously assumes a linear relationship between the latent happiness and the reported happiness scores. The boundedness of rating scales makes a linear relationship implausible.

Recent computer surveys that have implemented the LS and VAS experimentally used various methods to compare the rating scales. The item response times have been recorded. While Funke and Reips (2012) found no difference, Cook et al. (2001) and Couper et al. (2006) have reported a longer response time for the VAS. Completion rates of questionnaires have been lower and questions were skipped more often if the VAS instead of the LS was used (Couper et al., 2006). Answers were modified nearly twice as often with the VAS (Funke and Reips, 2012). Not all these indicators of response behaviors are measurable in this survey.

In the experiment, randomization took place only once the participants accessed the questionnaire. Therefore, item non-response cannot be assessed. Moreover, all participants

finished the questionnaire and completion rates do not differ. This study finds higher average item response times for the VAS (16 seconds) than for the LS (10 seconds). However, this may be due to a difference in question design: the VAS question had one sentence more to read (Figure 2.2). A higher fraction of survey participants was found to move back to adjust the happiness score for the VAS (2.3%) than for the LS (1.4%). This may simply indicate a lack of familiarity with the VAS as opposed to the LS.

## Validity and reliability

Validity and reliability are the most established measures to assess survey data quality. Validity quantifies the degree to which the rating scale is able to capture the true latent construct. A systematic error due to a nonconformity of a rating scale harms validity. Intuitively, the LS, requesting the categorization of a continuous feeling, may have lower validity than the VAS. Reliability is the extent to which the rating scale reproduces its measurements. Low reliability is due to a random measurement error. The high sensitivity of the VAS may lead to lower reliability. Different methodologies have been established to investigate validity and reliability of rating scales when the underlying dimension is latent.

The presence of validity in single-item happiness responses is evaluated through content or external validity (Diener, 1994). Content validity is assessed by the correlation between individual happiness scores of different rating scales. In this article the Spearman's rank correlation coefficient is used, instead of the Pearson's correlation coefficient. The former allows for a non-linear transformation function between VAS and LS scores. Only marginal differences in Spearman's rank correlation coefficients between the three scales are observed. First, the ranked VAS scores correlate with the ranked Likert 11 and 10 points scores by 0.68, whereas the rank correlation coefficient for the two discrete measurements is 0.71 (Table 2.3). Magnitudes of these point estimates are in line with earlier findings (e.g., Larsen et al., 1984). The positive correlations indicate that all three measures assess the same latent construct, but it cannot be concluded which scale is best.

External validity in contrast ranks scales. In happiness research the magnitudes of correlations between happiness scores and personality traits have been used. The analysis relies implicitly on the assumption of a valid scale of the external criterions, which are in most cases multiple item LS questions. Findings may thus be positively biased towards the LS, if LS in general induce systematic answer distortions. Estimates of Spearman's rank correlations between rating scales and the BIG FIVE inventory (Goldberger, 1992) or the self-esteem scale (Rosenberg, 1965) are reported in Table 2.4. These six trait variables were gathered in the Personality Study of the May wave. Personality traits have been reported to be stable (e.g., Srivastava et al., 2003) thus the time gap between the assessments of happiness (March, April, May) and personality traits (May) should not dilute estimates. The magnitudes of correlations are similar to earlier research (e.g., Larsen et al., 1984 or Abdel-Khalek, 2006) and no pattern in lower or higher external validity for one scale is apparent. Thus, based on the evidence presented in Table 2.4, the VAS appears to be equally valid to the 10 and 11 points LS measures of happiness.

The reliability of single-item happiness questions has been assessed through test-retest reliability. The test-retest method uses the same sample and the same measurement on two occasions. Larsen et al. (1984) and Krueger and Schkade (2008) have concluded on test-retest reliability coefficients ranging from 0.4 to 0.6 for single-item discrete measurements. The data structure of this study does not allow to present test-retest reliability coefficients.

However, reliability can be exploited using the experimental set-up of this study. It was shown earlier that randomization was successful and that no time effect exists. Sample distributions of happiness scores should equally map the latent happiness distributions in the March and April waves for each rating scale, if the scales are reliable. Figure 2.4 shows the histogram for the LS scores and the kernel density for the VAS scores, respectively. Substantial agreement between the March and April waves in the distributions of scores is observed for both scales. The equality of distributions among waves was tested for both random samples. A Pearson's Chi-Squared test for the discrete distributions (p-value=0.48)

and a Kolmogorov-Smirnov test for the continuous distributions (p-value=0.99) cannot reject the null hypothesis. The VAS is considered to be a reliable rating scale for happiness.

LS survey happiness data are widely accepted to be of good quality, e.g. to be valid and reliable measures. The existing methods to assess happiness data quality suggest no substantial differences between the VAS and the LS. Moreover, the theoretical argument is emphasized again: Higher (theoretical) validity should be attributed to the VAS, because it overcomes idiosyncratic discretization and the line length acts as a reference continuum to represent the underlying true happiness.

## 2.4 The distributions of happiness scores by scales

The random assignment of response scales creates two groups with the same latent distributions of true happiness. If both scales exhibited the same distributions in scores, question design effects would be absent. This section reports tests of equality of moments for distributions of reported happiness that exploit the experimental setup. Robustness of the findings is also investigated.

Existing experiments implementing LS and VAS focus on the first moment of score distributions. For instance, computer based studies have reported equal mean scores for the VAS and LS (e.g., Couper et al. 2006 or Funke and Reips, 2012) and a paper and pencil study reported lower mean values for the VAS (Flynn et al., 2006). However, equality in means does not imply equality in distributions.

The following thought experiment provides a good starting point. How has the VAS to be partitioned in 10 categories in order to replicate the LS distribution? For instance, equally spaced intervals of a length of 11 VAS points would suggest a linear transformation function between the two scales. Figure 2.5 shows the kernel estimate of the paired sample VAS distribution and the estimated cutoff values, that have to be chosen in order to replicate the LS distribution, in vertical lines. No systematic transformation function is apparent.

Substantial differences in interval widths are found. Intervals are narrow especially in the extreme categories, whereas intervals are wider for the discrete values 7 and 8. However, it is not apparent what these differences imply.

In order to understand the disagreement in distributions presented in Figure 2.5, statistical moments are compared. To do so, both scores were constrained to the closed interval ranging from 0 to 9, i.e. VAS scores were divided by 11. Table 2.5 reports the first and second moments of both happiness scales, the LS and the VAS. t-tests on the equality of means and Levine's tests on the equality of variances for each wave and for the paired sample are also presented. All three samples show the same picture: the VAS scores exhibit lower means but wider spread happiness scores. All null hypotheses of equality of means and variances can be rejected.

Are these disagreements in moments caused by the graphical design? In fact, the difference in mean happiness of 0.45 points may result from interpreting the boundaries of each scale differently. For instance, VAS scores would be artificially lower, if a LS score of 9 maps an interval of latent happiness ranging from 8.5 to 9.4, but a VAS score of 9 represents a latent happiness of 9 only. Furthermore, the second finding reported in Table 2.5, an increase in variances of 0.8 points, may be due to the high sensitivity of the VAS. For instance, people would like to mark with a cross the equivalent of a 7 but crossed 6.8 instead. In order to test these two hypotheses, VAS scores were transformed into discrete scores and the moments compared again. To discretize the VAS, the line was divided into ten equally spaced intervals. The intervals were assigned the LS scores 0 to 9 in ascending order from left to right. The difference in means decreases by 0.2 points, but remains negative and significant. The variance increases by another 0.5 points. This evidence suggests differences in the first two moments caused by question design.

It is conceivable that these question design effects can be explained by a latent rescaling of scales. Score distributions can differ among scales, while each scale's distribution still represents the same ranking of individuals. In order to abstract from scaling effects, an

equal unit of measurement for both scales has to be established. This can be done by setting the means to 0 and the variances to unity for each scale. The resulting dimensionless measures have as unit of measurement one standard deviation and indicate each individual's relative position compared to the mean and the variation of the observed distribution. If different scaling is the reason for question design effects, no differences in the distributions of standardized scores should exist.

In order to analyze differences in standardized scores, a variable indicating the location of individual standardized scores was generated. The variable distinguishes three states. It takes the value one if the individual answered below -1 standard deviation, the value two if the participant reported in between the range of -1 and 1 standard deviation and the value three if the answer was more than 1 standard deviation above the mean. The variable is modeled by the multinomial logit model. The scores close to the mean, i.e. the indicator variable is equal to two, form the baseline category. A large set of socioeconomic and sociodemographic variables as well as a wave dummy and dummies indicating the questionnaire and question order are included in the regression. Table 2.6 reports the estimated odds ratios. The odds ratios are equal to the factor of increase in the probability of an extreme score relative to the probability of a score close to the mean. In fact, the odds for positive and negative extreme scores relative to non-extreme answers increase significantly by two if the VAS instead of the LS was used. Standardized distributions differ between the two scales. Differences in distributions are not attributable to rescaling mechanisms. The VAS leads to wider spread answers.

The analyses reveal robust question design effects. The two rating scales cause different happiness score distributions. An increased likelihood of more extreme answers on the VAS is observed. This finding provides evidence that the VAS contains more information than the LS. More than 70% of all participants answered a 7 or 8 on the LS. The VAS demonstrates that there is more variation in happiness in the population than the LS suggests. The LS high frequency categories 7 and 8 seem to be a scale artifact. In fact,

an earlier international comparison concluded on Dutch people being more likely to avoid extreme LS values (Kapteyn et al., 2007). The present results show that Dutch respondents are willing to score closer to the boundaries, but maybe not at the boundaries itself. A measurement with continuous categories enables respondents to approach the boundaries and thus overcomes an endpoint aversion present in a too insensitive answer scale, as the LS seems to be.

## 2.5 The correlates of happiness by scales

Research into the determinants of subjective well-being has burgeoned in recent years, and valuable insights have been obtained (e.g., Kahneman and Krueger, 2006). Scholars have been interested in the effects of schooling (Orepolus, 2003), income (Easterlin, 1995), unemployment (Winkelmann and Winkelmann, 1998) or age (Stone et al., 2010). Many findings have been replicated for different countries and have been judged as robust (Frey and Stutzer, 2002). All these studies use discrete happiness data. Therefore, the question arises: How much are these findings affected by the specificities of the LS?

The paired sample consists of the same set of respondents assessing their happiness on the VAS and before or after also on the LS. From March to April, no individual reported changes in core socioeconomic or sociodemographic variables. Regressions for both scales of standardized happiness scores on a set of socioeconomic and sociodemographic variables should estimate the same effects. Any changes in correlates when moving from one to the other scale can be attributed to the change in scale design.

Table 2.7 shows estimates by scale types of regressions modeling standardized happiness scores as dependent variable. Again the Rating Scale Model employing a logit-type link function and a Bernoulli quasi-maximum likelihood estimation is used to account for the boundedness of the measurements. Average marginal and average discrete effects for the LS are in line with the research literature (e.g. Kahneman and Krueger, 2006 or Frey

and Stutzer, 2002). Happiness is found to be increasing in income and U-shaped in age. Foreigners are less happy and women and employed are more happy. House ownership, which may be interpreted as a proxy for savings, as well as marriage and cohabitation with a partner have a positive effect on happiness.

A comparison of correlation coefficients in Table 2.7 reveals some striking differences. Signs of statistically significant explanatory variables stay the same for both scales. Also, except for the male dummy, effects of statistically significant variables are similar in magnitudes. Three variables lose significance in the VAS regression, the male and questionnaire order dummies as well as the number of household members. Whereas the estimates for the two latter variables are similar in magnitudes in both regressions, the substantial gender differences in average LS scores vanishes completely once the VAS scores are compared. This finding needs some investigation.

Men are found to be 0.13 standard deviations less happy than women in the LS data. This is a substantial gender inequality. For instance, the compensating income variation, i.e. the income increase necessary to make a man as happy as a comparable woman, is estimated to 267%. This effect is in line with evidence presented in crossectional studies using Likert type single-item happiness questions as dependent variable (e.g., Wood et al., 1989; Gerdtham and Johannesson, 2001; Lalive and Stutzer, 2010). Especially the psychological literature has been focusing on such gender differences in well-being and provided several explanations (for a review see Nolen-Hoeksema and Rusting, 1999). However, by use of the VAS to assess happiness, the happiness gender inequality vanishes completely.

The differences in correlates indicate that subgroups of the population are influenced to different degrees by rating scale design. Gender is found to play a major role in perception of answer scales. However, it remains to show if male report higher or women lower happiness on the VAS. To do so a regression of the indicator variable identifying extreme answers is run on the set of explanatory variables, question type and gender dummies and an interaction term between the two dummy variables. Again a multinomial logit model is

employed with the base category being the scores centered around the mean. Table 2.8 reports the estimated odds ratios for the use of the VAS by gender. The VAS dummy reveals an important finding. The odds of extreme answers relative to centered answers are found to more than double, if female participants used the VAS. Moreover, women's relative probability of a large negative score increases more than the relative probability of a large positive score once the VAS is used. For men the VAS causes less variation in scores. The odds of negative scores compared to centered scores decrease significantly, by marginal 4 percentages. In a nutshell, women are reacting stronger to the VAS than men. Women are likelier to score towards the endpoints and to revise their answers downwards on the VAS than men are. The disappearance of the gender happiness gap is mainly explainable by the answer style of female participants. This finding casts on reliability of inferences on the gender gap by earlier studies.

What drives the disappearance of the gender happiness inequality? A possible explanation is that respondents think about their happiness in numbers and have to find an equivalent category to the number in mind on the VAS. In such a scenario the disappearance of the gender happiness gap reflects problems women face in attributing numbers to intervals without any visual heuristics. If this hypothesis held true, women should be expected to manifest response difficulties. Table 2.9 reports summary statistic that should detect women's answer difficulties, if any of those were present. For men and women, who have responded the VAS, mean values for answer times measured in seconds and for the evaluations on question clearness and difficulty are shown. Moreover, the proportion of individuals having updated the Background Variables questionnaire during the two hours preceding the happiness questionnaire are reported. An equal proportion of women and men responded to the happiness questionnaire after having had answered other questions. Possible framing with discrete question types is not more likely for one or the other gender. Moreover, women do not manifest problems in responding to the VAS. They respond as fast as male respondents and judge the overall question equally clear and difficult as men

do. None of the differences in means is significant. There exists no evidence that women have problems finding an equivalent of a number on the VAS. This suggests that women and men use the VAS similarly.

The disappearance of the gender happiness gap is caused by focal points shifting differently among scales for women and men. This study's experiment is unable to fully attribute it to one or the other scale. However, support for the hypothesis that a gender gap in happiness results from numerically labeled LS can be found. First, the May wave of the data provides a second discrete measurement of happiness. The regression reported in Table 2.7 was repeated using the LS data with 11 categories ranging from 0 to 10. Estimates, not reported in this paper, again identified women to be nearly one tenth of a standard deviation happier than men on average. Second, it is hard to think of any plausible economic explanation why coequal men or women should derive different satisfaction levels given the same socioeconomic and sociodemographic backgrounds. Third, the psychological literature dedicated to question design generally reports strong dependence of answers on numerical labels (e.g., Schwarz, 1991; Dillman, 2002). Last, this article adds evidence to a recent article by Conti and Pudney (2011) in which the authors report changes in women's job satisfaction levels fully attributable to numerical labeling. Discrete numerically labeled single-item happiness questions seem to be responsible for the differences in happiness between gender found earlier in the literature.

## 2.6 Conclusions

Most of the studies concerning the determinants of happiness have used discrete satisfaction scores as dependent variables. Such Likert scale measures are widely available in crossectional or panel surveys. The evidence presented in this article suggests that there are gains in moving away from the discrete Likert scale. The visual analogue scale, a continuous measurement, was implemented in the Dutch Longitudinal Internet Study for

Social Sciences. This study is the first to exploit a randomized controlled experiment to compare a single-item happiness question assessed either on a LS or on a VAS. Results are promising. First, survey participants did not manifest problems in using the VAS. Second, no differences in data quality were found between the VAS and LS. Third, the VAS scores provide more information than the LS scores. The use of the VAS increases the likelihood of participants scoring close to the boundaries. This finding reveals the high frequency LS categories 7 and 8 as a result of too little discriminating power. Fourth, gender specific question design effects were found. Female participants' reports using the LS differ significantly from continuous counterparts. Men reacted less to question design. It is likely that this gender specific response behavior is at the heart of the gender happiness inequality, i.e. women being on average happier than men, which has been reported earlier. In a nutshell, differences between both scales are found. The empirical findings, but also the theoretical argument, favor the VAS. The underlying dimension, happiness, is continuous. Hence, the VAS can be interpreted as a reference continuum.

Both scales are found to be consistent in their way of assessing happiness. But the VAS interpretable as a reference continuum for the latent continuous happiness overcomes discretization and ordinal measurement of well-being. Furthermore, analyses suggest that the VAS is preferable to the LS. The VAS provides more information by overcoming the endpoint aversion observed with the LS. These results hopefully encourage researchers to employ computer-based survey technology in order to develop new subjective well-being measurements that contribute leading happiness economics to its next level.

# References


Abdel-Khalek, A. M., 2006, "Measuring Happiness with a Single-Item Scale", *Social Behavior and Personality*, 34, 2, 139-150

Andrews F. M. and R. Crandall, 1976, "The Validity of Measures of Self-Reported Well-Being", *Social Indicators Research*, 3, 1-19

Bouazzaoui, A. B. and E. Mullet, 2006, "Employment and Family as Determinants of anticipated Life Satisfaction: Contrasting European and Maghrebi People's Viewpoints", *Journal of Happiness Studies*, 6, 161-185

Conti, G. and S. Pudney, 2011, "Survey Design and the Analysis of Satisfaction", *The Review of Economics and Statistics*, 93, 3, 1087-1093

Cook, C., F. Heath and R. L. Thompson, 2001, " Score reliability in web- or Internet-based surveys: Unnumbered graphic rating scales versus Likert-type scales", *Educational and Psychological Measurement*, 61, 697-706

Couper, M. P., R. Tourangeau, F. G. Conrad and E. Singer, 2006, "Evaluating the Effectiveness of Visual Analog Scales : A Web Experiment", *Social Science Computer Review*, 24, 227-245

Cummins, R. A., 2003, "Normative Life Satisfaction: Measurement Issues and a Homeostatic Model", *Social Indicators Research*, 64, 225-256

de Vos, K., 2010, "Representativeness of the LISS-panel 2008, 2009, 2010", published online http://www.lissdata.nl, last consultation 14.10.2011

Diener, E., 1994, "Assessing Subjective Well-Being: Progress and Opportunities", *Social Indicators Research*, 31, 2, 103-157

Dillman, D. A., 2007, "Mail and Internet Surveys. The Tailored Design Method.", New Jersey: Hoboken

Dillman, D. A. and L. M. Christian, 2002, "The Influence of Words, Symbols, Numbers, and Graphics on Answers to Self-Administered Questionnaires: Results from 18 Experimental Comparisons", published online http://sesrc.wsu.edu/dillman/papers/2002/theinfluencewords.pdf, last consultation 08.08.2012

Easterlin, R., 1995,"Will Raising the Incomes of All Increase the Happiness of All?", *Journal of Economic Behavior and Organization*, 27, 1, 35-48

Flynn, D., P. van Schaik and A. van Wersch, 2004, "A Comparison of Multi-Item Likert and Visual Analogue Scales for the Assessment of Transactionally Defined Coping Function", *European Journal of Psychological Assessment*, 20, 1, 49-58

Fordyce, M. W., 1987, "A Review of Research on the Happiness Measures: A Sixty Second Index of Happiness and Mental Health", In Alex C. Michalos (Ed.), *Citation Classics from Social Indicators Research*, 2005, 373-399

Frey, B. S. and A. Stutzer, 2002, "Happiness and Economics", Princeton, NJ: Princeton University Press

Funke, F., U.-D. Reips and R. K. Thomas, 2010, "Sliders for the Smart: Type of Rating Scale on the Web Interacts With Educational Level", *Social Science Computer Review* published online August 16 2010, DOI 10.1177/0894439310376896

Funke, F. and U.-D. Reips, 2012, "Why Semantic Differentials in Web-Based Research Should be Made From Visual Analogue Scales and Not From 5-Point Scales", *Field Methods*, 24, 310-324

Gerdtham, U. G. and M. Johannesson, 2001, "The relationship between happiness, health, and socioeconomic factors: results based on Swedish microdata", *Journal of Socio-Economics*, 30, 553-557

Goldberger, L. R., 1992, "The Development of Markers for the Big-Five Factor Structure", *Psychological Assessment*, 4, 3, 26-42

Hayes, M. H. S. and D. G. Patterson, 1921, "Experimental development of the graphic rating method", *Psychological Bulletin*, 18, 98-99

Hofmans, J. and P. Theuns, 2008, "On the linearity of predefined and self-anchoring Visual Analogue Scales", *British Journal of Mathematical and Statistical Psychology*, 61, 401-413

Kahneman D., E. Diener, and N. Schwarz (Eds.), 1999, *Well-Being: The Foundations of Hedonic Psychology*, New York: Russell Sage Foundation

Kahneman, D. and A. B. Krueger, 2006, "Developments in the Measurement of Subjective Well-Being", *Journal of Economic Perspectives*, 20, 1, 3-24

Kapteyn, A., J. P. Smith and A. van Soest, 2007, "Vignettes and self-reports of work disability in the U.S. and the Netherlands", *American Economic Review*, 97, 461-473

Knoef M. and K. de Vos, 2009, "The representativeness of LISS, an online probability panel", published online http://www.lissdata.nl, last consultation 14.10.2011

Kreindler, D., A. Levitta, N. Woolridge and C.J. Lumsdenc, 2003, "Portable moodmapping: the validity and reliability of analog scale displays for mood assessment via hand-held computer", *Psychiatry Research*, 120, 165-177

Kristoffersen, I., 2010, "The Metrics of Subjective Wellbeing: Cardinality, Neutrality and Additivity", *The Economic Record*, 86, 272, 98-123

Krueger, B. and D. A. Schkade, "The Reliability of Subjective Well-Being Measures", *Journal of Public Economics*, 92, 1833-1845

Lalive R. and A. Stutzer, 2010, "Approval of equal rights and gender differences in well-being", *Journal of Population Economics*, 23, 933-962

Lara-Muñoz, C., S. P. de Leon, A. R. Feinstein, A. Puente and C. K. Wells, 2004, "Comparison of Three Rating Scales for Measuring Subjective Phenomena in Clinical Research. I. Use of Experimentally Controlled Auditory Stimuli", *Archives of Medical Research*, 35, 43-48

Larsen, R. J., E. Diener and R. A. Emmons, 1985, "An Evaluation of Subjective Well-Being Measures",*Social Indicators Research*, 17, 1, 1-17

Layard, R., S. Nickell and G. Mayraz, 2008, "The marginal utility of income", *Journal of Public Economics*, 92, 1846-1857

Likert, R., 1932, "A Technique for the Measurement of Attitudes", *Archives of Psychology*, 140, 1-55

Matsubayashi K., S. Kimura, T. Iwasaki, K. Okumiya, T. Hamada, M. Fujisawa, K. Takeuchi, T. Kawamoto and T. Ozawa, 1992, "Application of visual analogue scale of happiness to elderly Himalayan highlanders", *Nippon Ronen Igakkai Zasshi*, 29, 11, 823-828

McCormack, H. M., D. J. L. Horne and S. Sheater, 1988, "Clinical Applications of Visual Analogue Scales: A critical Review", *Psychological Medicine*, 18, 1007-1019

Nolen-Hoeksema, S. and C. L. Rusting, 1999, "Gender Differences in Well-Being", In D. Kahneman, E. Diener and N. Schwarz (Eds.), *Well-Being: The Foundations of Hedonic Psychology*, New York: Russell Sage Foundation

Oreopoulos, P., 2003, "Do Dropouts Drop Out Too Soon? Evidence from Changes in School-Leaving Laws", mimeo, University of Toronto, March

Paul-Dauphin, A., F. Guillemin, J. M. Virion and S. Briancon, 1999, "Bias and Precision in Visual Analogue Scales: A Randomized Controlled Trial", *American Journal of Epidemiology*, 150, 10, 1117-1127

Rosenberg, M., 1965, "Society and the adolescent self-image", Princeton, NJ: Princeton University Press

Saris, W. E., T. van Wijk and A. Scherpenzeel, 1998, "Validity and Reliability of Subjective Social Indicators", *Social Indicators Research*, 45, 173-199

Saris, W. E. and I. N. Gallhofer, 2007, "Design, Evaluation, and Analysis of Questionnaires for Survey Research", New Jersey: Hoboken

Scherpenzeel, A., "Start of the LISS panel: Sample and recruitment of a probability-based Internet panel", pusblished online http://www.lissdata.nl, last consultation 14.10.2011

Schuman, H. and S. Presser, 1981, "Questions and Answers in Attitudes Surveys: Experiments on Question Form, Wording and Context", New York: Academic Press

Schwarz, N., B. Knäuper, H.-J. Hippler, E. Noelle-Neumann and L. Clark, 1991, "Rating Scales: Numeric Values May Change the Meaning of Scale Labels", *Public Opinion Quarterly*, 55, 570-582

Srivastava, S., O. P. John, S. D. Gosling and J. Potter, 2003, "Development of personality in early and middle adulthood: Set like plaster or persistent change?", *Journal of Personality and Social Psychology*, 84, 1041-1053

Stone, A. A., J. E. Schwartz, J. E. Brodericka and A. Deaton, 2010, "A Snapshot of the Age Distribution of Psychological Well-being in the United States", *PNAS Paper*, published online May 17 2010, DOI 10.1073/pnas.1003744107

Treiblmaier, H. and P. Filzmoser, 2011, "Benefits from using continuous rating scales in online survey research", *ICIS 2011 Proceedings*, Paper 1

van Praag, B. M. S. and A. Ferrer-i-Carbonell, 2004, *Happiness Quantified: A Satisfaction Calculus Approach*, New York: Oxford University Press

Weng, L.-J., 2004, "Impact of The Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability", *Educational and Psychological Measurement*, 64, 956-972

Winkelmann, L. and R. Winkelmann, 1998, "Why Are the Unemployed So Unhappy? Evidence from Panel Data", *Economica*, 65, 257, 1-15

Wood, W., N. Rhodes and M. Whelan, 1989, "Sex Differences in Positive Well-Being: A Consideration of Emotional Style and Marital Status", *Psychological Bulletin*, 106, 2, 249-264

Table 2.1: Test for randomization

| | LS | | VAS | | Mean equality |
|---|---|---|---|---|---|
| | Obs | Mean | Obs | Mean | t-test (p-value) |
| Proportion male | 2129 | 0.49 | 2039 | 0.50 | 0.36 |
| Net monthly income (EUR) | 2129 | 1716.27 | 2039 | 1651.65 | 0.58 |
| Age | 2129 | 51.35 | 2039 | 52.36 | 0.05 |
| Number of hh-members | 2129 | 2.52 | 2039 | 2.47 | 0.19 |
| Proportion house owner | 2129 | 0.73 | 2039 | 0.72 | 0.20 |
| Proportion unemployed | 2129 | 0.03 | 2039 | 0.03 | 0.82 |
| Proportion working | 2129 | 0.58 | 2039 | 0.55 | 0.08 |
| Proportion secondary educated | 2129 | 0.35 | 2039 | 0.36 | 0.39 |
| Proportion tertiary educated | 2129 | 0.57 | 2039 | 0.55 | 0.43 |
| Proportion married | 2129 | 0.57 | 2039 | 0.61 | 0.04 |
| Proportion separated | 2129 | 0.10 | 2039 | 0.10 | 0.91 |
| Proportion foreigner | 2129 | 0.13 | 2039 | 0.10 | 0.00 |

*Notes:* March sample employed.

Table 2.2: Regressions of happiness on wave and questionnaire order dummies by scales

| | LS | VAS |
|---|---|---|
| April | -0.007 | -0.528 |
| | (0.040) | (0.572) |
| Experiment $2^{nd}$ | 0.013 | -1.374** |
| | (0.048) | (0.685) |

*Notes:* $N_{ls} = 4122$ and $N_{vas} = 4028$; Rating Scale Model estimates of average discrete effects are shown. Experiment $2^{nd}$ equals 1 if during the 2 hours preceding the happiness questionnaire the background variable questionnaire was answered. Heteroscedasticity consistent standard errors presented in parentheses. ***, **, * denote statistical significance at the 1%, 5%, 10% significance levels, respectively.

Table 2.3: Convergent validity of happiness scales - Spearman's rank correlations

| | VAS March/April | LS March/April | LS May |
|---|---|---|---|
| VAS March/April | 1 | 0.68 | 0.68 |
| LS March/April | | 1 | 0.71 |
| LS May | | | 1 |

*Notes:* $N = 3987$; The LS of the March and April waves ranged from 0 to 9, whereas the LS of the May wave from 0 to 10.

Table 2.4: External validity of happiness scales - Spearman's rank correlations

| | VAS March/April | LS March/April | LS May |
|---|---|---|---|
| Extraversion | 0.19 | 0.19 | 0.20 |
| Agreeableness | 0.09 | 0.11 | 0.12 |
| Consciousness | 0.18 | 0.18 | 0.20 |
| Emotional stability | 0.40 | 0.38 | 0.41 |
| Openness to experience | 0.02 | 0.02 | 0.05 |
| Self-esteem | 0.38 | 0.37 | 0.41 |

*Notes:* $N = 3987$; The LS of the March and April waves ranged from 0 to 9, whereas the LS of the May wave from 0 to 10. Spearman's rank correlation coefficients are presented.

Table 2.5: Mean and standard deviations of happiness scores by scales and waves

| | LS | VAS | Mean equality t-test | Variance equality Levine's test |
|---|---|---|---|---|
| March wave | 7.15 | 6.70 | 0.00 | |
| | (1.22) | (1.52) | | 0.00 |
| April wave | 7.14 | 6.70 | 0.00 | |
| | (1.17) | (1.50) | | 0.00 |
| Paired sample | 7.15 | 6.70 | 0.00 | |
| | (1.19) | (1.51) | | 0.00 |

*Notes:* Unpaired sample sizes: $N_{ls,march} = 2129$, $N_{ls,april} = 1993$, $N_{vas,march} = 2039$, $N_{vas,april} = 1989$; paired sample sizes: $N_{ls,march} = 1788$, $N_{ls,april} = 1769$; Standard deviations presented in parentheses. P-values reported for tests.

Table 2.6: Differences in distributions of standardized happiness scores between scales

| | Changes in the odds of extreme scores relative to the base category | |
|---|---|---|
| | scores smaller than $-\sigma$ | scores larger than $\sigma$ |
| VAS dummy | 1.971*** | 2.191*** |
| | (0.115) | (0.146) |

*Notes:* $N = 8150$; Odds ratios estimated by multinomial logit. The base category consists of standardized happiness scores in-between $-1$ and 1 standard deviations. Control variables include: time dummy, order of questionnaires, gender, log of income, age, age$^2$, log of number of hh members, cohabitation with partner, house ownership, employment and marital status, education level and origin. Heteroscedasticity consistent standard errors presented in parentheses. ***, **, * denote statistical significance at the 1%, 5%, 10% significance levels, respectively.

Table 2.7: Regressions of standardized happiness scores on characteristics by scales

| | LS | | VAS | |
|---|---|---|---|---|
| | Coefficient | S.E. | Coefficient | S.E. |
| Male | -0.131*** | 0.035 | -0.007 | 0.036 |
| Log of net monthly income (EUR) | 0.099*** | 0.030 | 0.105*** | 0.030 |
| Age | -0.037*** | 0.008 | -0.030*** | 0.007 |
| $\text{Age}^2 \cdot 10^{-2}$ | 0.006*** | 0.001 | 0.004*** | 0.001 |
| Log number of hh-members | -0.106** | 0.048 | -0.058 | 0.050 |
| Cohabiting | 0.269*** | 0.060 | 0.192*** | 0.062 |
| House ownership | 0.214*** | 0.040 | 0.227*** | 0.039 |
| In workforce | 0.091* | 0.049 | 0.105** | 0.049 |
| Unemployment | -0.092 | 0.103 | 0.026 | 0.108 |
| Secondary education | -0.025 | 0.068 | -0.013 | 0.069 |
| Tertiary education | 0.001 | 0.066 | -0.048 | 0.069 |
| Married | 0.293*** | 0.054 | 0.316*** | 0.054 |
| Separated | -0.026 | 0.062 | -0.067 | 0.062 |
| Foreigner | -0.135** | 0.053 | -0.113** | 0.052 |
| Experiment $2^{nd}$ | 0.070* | 0.041 | 0.020 | 0.043 |
| April dummy | 0.002 | 0.034 | 0.011 | 0.035 |

*Notes:* $N_{ls} = N_{vas} = 3557$; Rating Scale Model estimates of average marginal and discrete effects are shown. Heteroscedasticity consistent standard errors presented in parentheses. ***, **, * denote statistical significance at the 1%, 5%, 10% significance levels, respectively.

Table 2.8: A closer look at gender differences: Heterogeneous effects of scale design on standardized happiness

| | Changes in the odds of extreme scores relative to the base category | |
|---|---|---|
| | scores smaller than $-\sigma$ | scores larger than $\sigma$ |
| VAS · Female | 2.275*** | 2.103*** |
| | (0.187) | (0.215) |
| VAS · Male | 0.963*** | 1.079 |
| | (0.107) | (0.131) |

*Notes:* $N = 7114$; Odds ratios estimated by multinomial logit. The base category consists of standardized happiness scores in-between $-1$ and 1 standard deviations. Control variables include: time dummy, order of questionnaires, gender, log of income, age, age$^2$, log of number of hh members, cohabitation with partner, house ownership, employment and marital status, education level and origin. Heteroscedasticity consistent standard errors presented in parentheses. ***, **, * denote statistical significance at the 1%, 5%, 10% significance levels, respectively.

Table 2.9: A closer look at gender differences: Answer difficulties for the VAS

| | Male | Female |
|---|---|---|
| Experiment $2^{nd}$ | 0.23 | 0.23 |
| Clearness | 4.40 | 4.40 |
| Answer difficulty | 1.59 | 1.60 |
| Response time | 16.62 | 16.14 |

*Notes:* $N_{male} = 1776$, $N_{female} = 1780$; Experiment $2^{nd}$ is a dummy variable indicating whether the LISS panel Background Variables questionnaire was answered before the experiment. Clearness of the question and answer difficulty were assessed on a scale ranging from 1 to 5. Response time is measured in seconds. The hypotheses of equality of means cannot be rejected by a t-test at the 1% significance level for all variables.

Figure 2.1: Data structure: stocks and flows

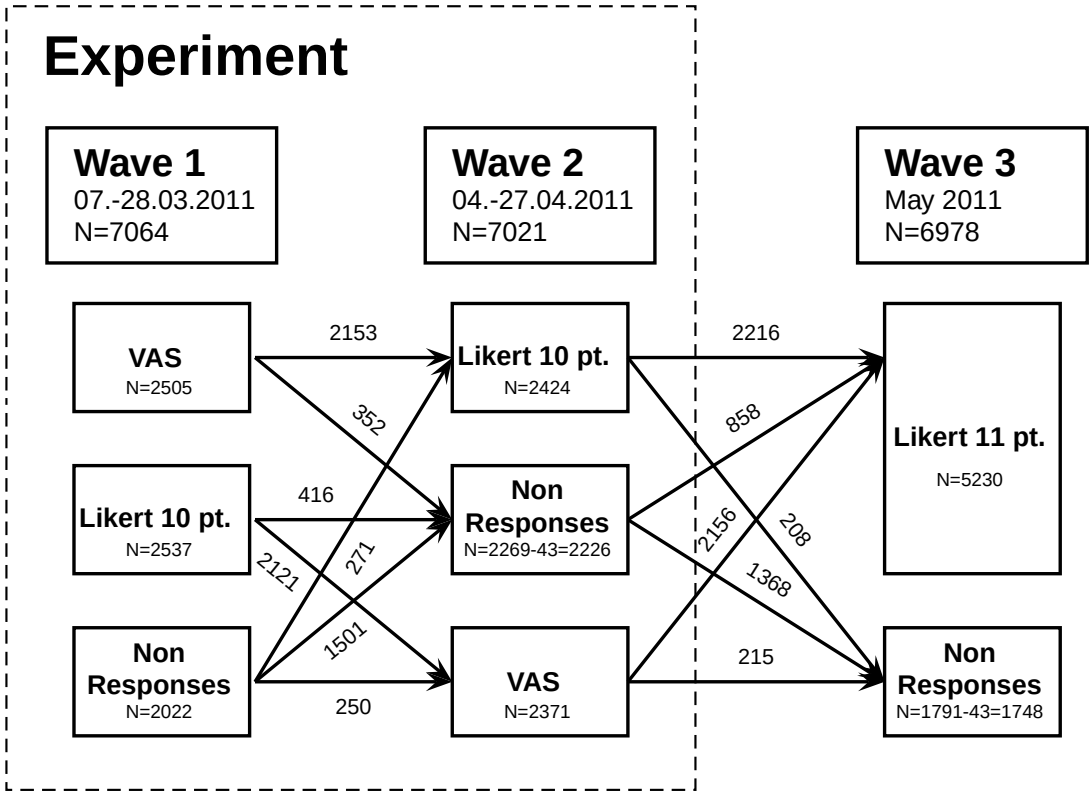


Figure 2.2: Screenshots of happiness questions

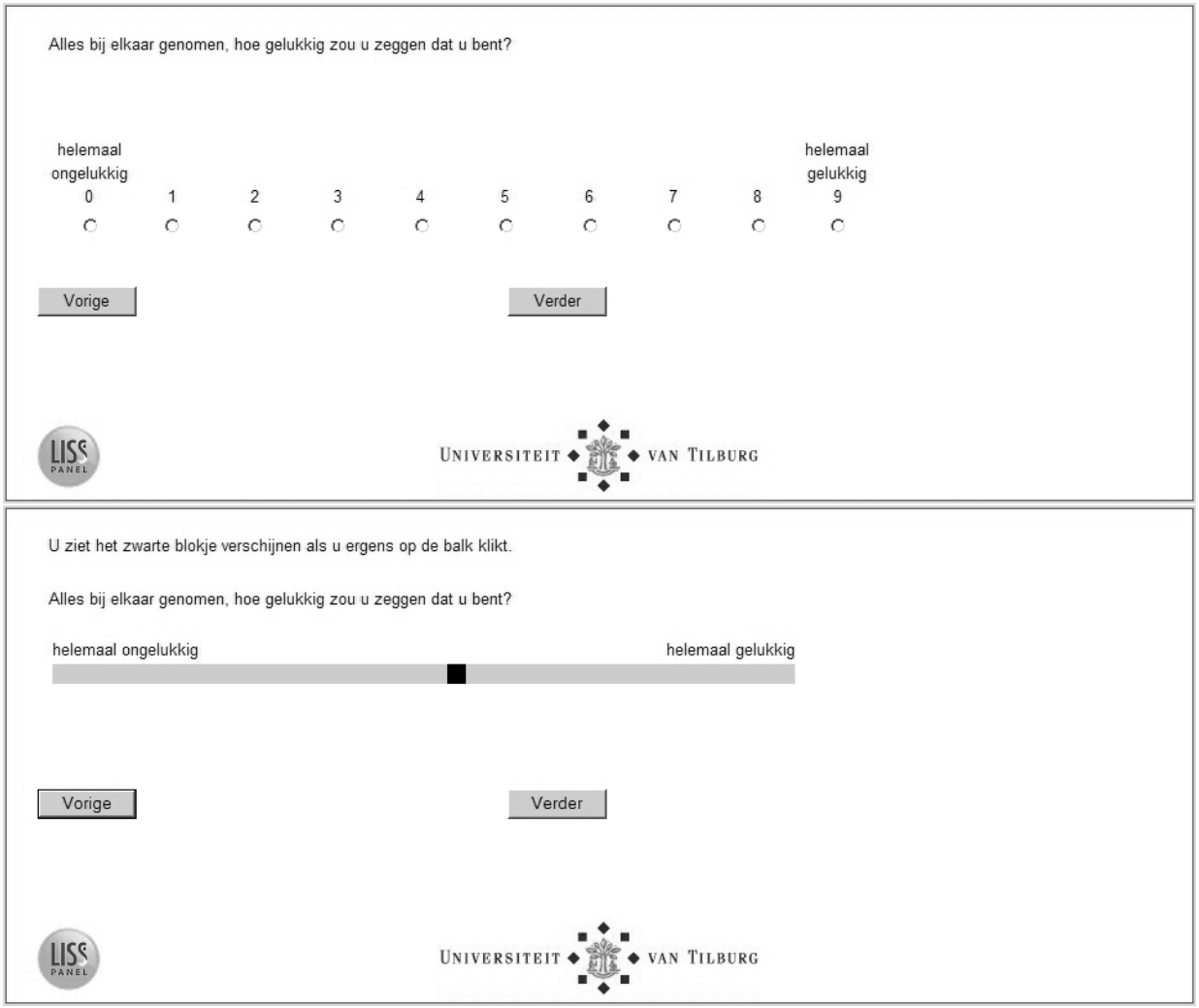
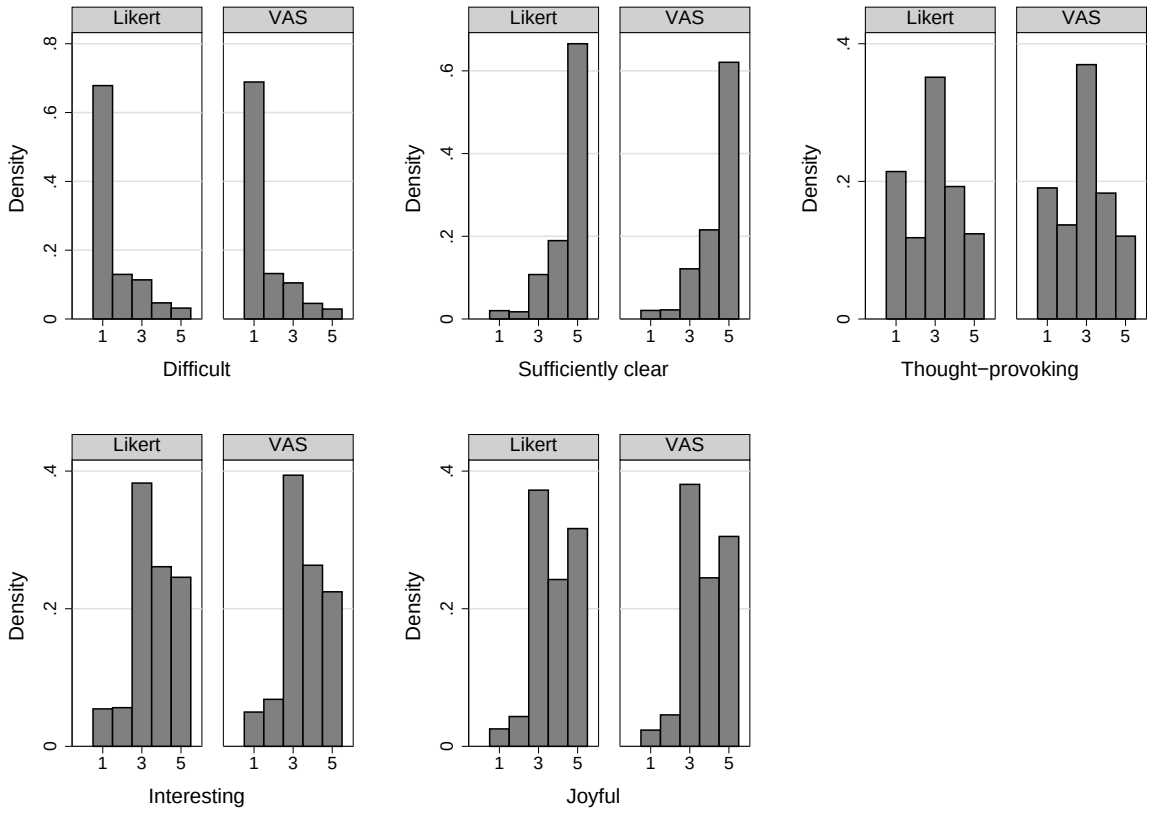
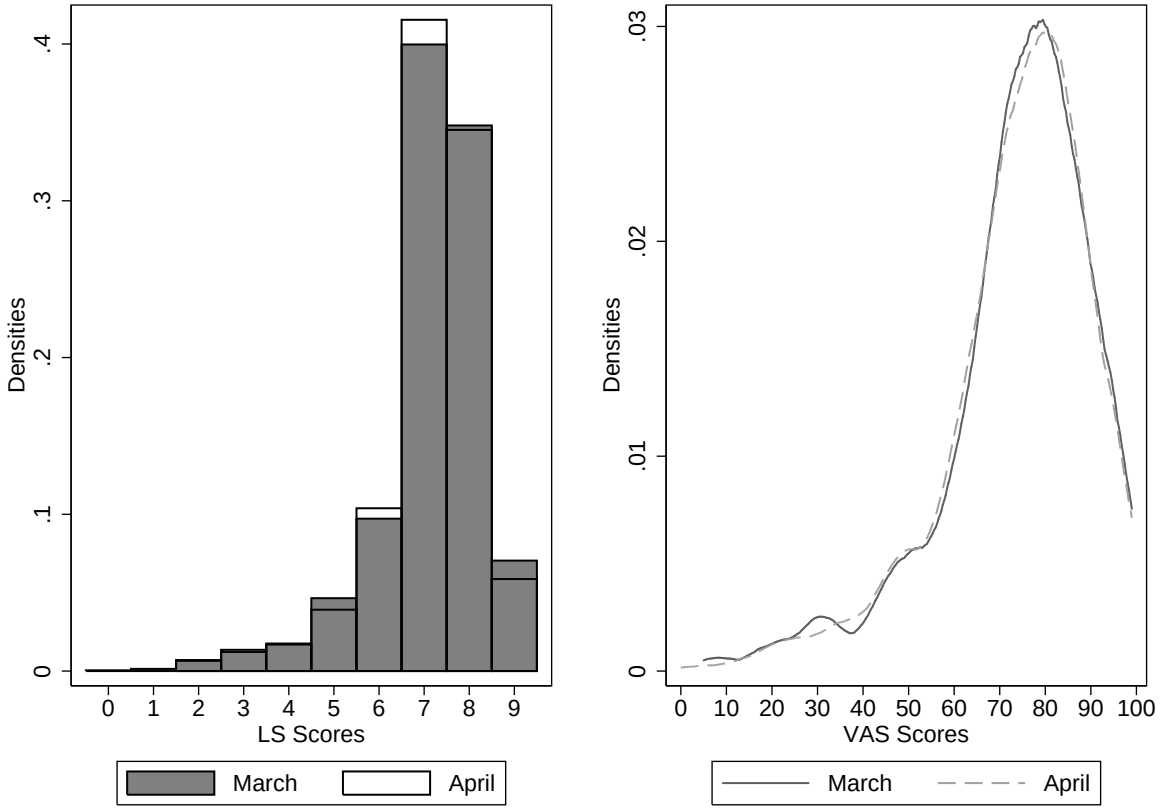


Figure 2.3: Densities of answers to questionnaire evaluation questions by scales



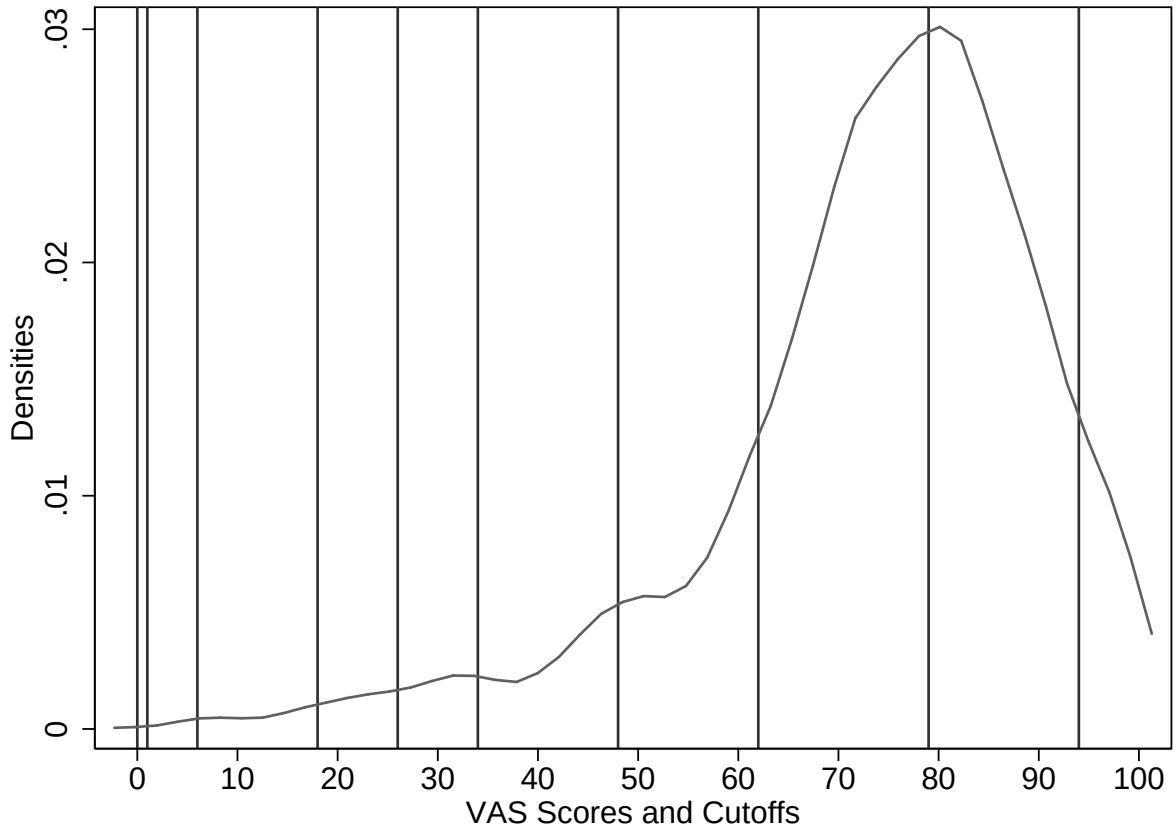

*Notes:* $N_{ls} = 4961$ and $N_{vas} = 4875$; The evaluation questions appeared as a cluster on the screen after the particular happiness question was answered.

Figure 2.4: Happiness densities for March and April waves by scales




*Notes:* $N_{ls,march} = 2129$, $N_{ls,april} = 1993$, $N_{vas,march} = 2039$ and $N_{vas,april} = 1989$; An Epanechnikow kernel with bandwidths equal to 2.6 respectively 2.7 is used to estimate VAS March and April densities.

Figure 2.5: Transformation function of continuous scores to discrete scores




*Notes:* $N_{ls} = 4122$ and $N_{vas} = 4028$; Vertical lines show how VAS scores need to be grouped in order to replicate the distribution of LS scores. An Epanechnikow kernel with bandwidth equal to 2.3 is used to estimate VAS densities.

# Chapter 3

# Reported happiness, fast and slow

This chapter is co-authored with Rainer Winkelmann. An earlier version was published as Working Paper No. 80 in the *Working Paper Series* of the Department of Economics, University of Zurich.

*Acknowledgements*: The paper owes many insights, as well as the title, to Kahneman (2011). We are grateful to participants at the 2012 EEA annual congress, the MESS workshop and the HUI and Sinergia seminars of the University of Zurich for helpful comments and suggestions, in particular to Bruno Frey, Arie Kapteyn, Bert van Landeghem, Arthur Stone and Josef Zweimüller. The paper uses data of the 2011 wave of the LISS panel of CentERdata.

## 3.1 Introduction

A standard definition of happiness is the degree to which one evaluates one's life-as-a-whole positively (Diener, 1984, Veenhoven, 1984). It emphasizes the subjective nature of happiness, and it suggests that self-reported data, perhaps based on a single question such as "All things considered, how happy would you say you are?", can provide direct and valid information on the individual's state of happiness. A question of this type has been included in many surveys and answered by countless individuals. Yet there is still some skepticism as to whether this measurement provides useful information. First, people may answer without giving it much thought, or even willfully misstating their happiness. Second, people may try to provide an honest evaluation of their stable inner state of well-being but use heuristics and thought patterns that lead to a dominance of transient effects of situational variables.

The first objection is of course common to any survey answer, and not specific to happiness. Item non-response rates on happiness questions are very low, and people seem not find it hard to answer. The second objection has been addressed in either one of two ways by the previous literature. The first approach has been to generate experimental variation in salience or mood at the time of the survey, and then study the effect on happiness responses. In a by now classical study, Schwarz (1987) reported that people who were put in a good mood by finding a coin placed by the experimenter reported a markedly higher life satisfaction than those who didn't. Schwarz and Clore (1983) exploited natural variation in weather to show that participants reported higher life satisfaction on sunny days than on rainy days. Strack, Martin and Schwarz (1988) randomly varied a contextual factor, namely the item order within a questionnaire. They found that the responses to questions on dating frequency and life satisfaction correlated strongly, when the life satisfaction question came second, and practically not at all, when life satisfaction was asked first.

The second approach uses the idea that any sensitivity to transient influences should

be reflected in a low test-retest stability. The evidence is mixed. Although Krueger and Schkade (2008) conclude that reported well-being changes as much over the short run as affect measures, Pavot and Diener (1993) had documented a significant degree of stability in self-reported happiness over time in an earlier study.

In this paper, we introduce a third, different approach, to further probe into the relationship between self-reported happiness and the underlying long-term, stable happiness. We do this by *measuring* contextual variables and including them as regressors in happiness equations. The three variables "response time" , "questionnaire order", and "self-assessed cognitive effort" capture aspects of the answering process. Since they are chosen by the respondent rather than being exogenously determined, we cannot claim a causal relationship. Nevertheless, we argue below that analyzing these "reporting correlates" can provide useful information on the relationship between reported happiness and underlying long-term happiness.

Our data come from a large representative household survey in the Netherlands, with more than 4000 respondents. Internet-based survey technology allows us to capture aspects of the response process in a non-intrusive way, based purely on technical information on the data flow between the data server and the personal computer of the respondent. Response time is simply the time between display of a question and entering of the response. Within a given session, some respondents are shown a link to more than one questionnaire, and we know the order in which they were opened (the order of questions within a questionnaire cannot be affected). Finally, participants are asked to what extent the happiness question "made them think". While this is a conventional question unrelated to technology, we include it in our analysis since it also relates to the way respondents answer the happiness question. To the best of our knowledge, this is the first time that such information is used in the context of happiness equations. Our general approach is similar to Rubinstein (2007) who used online response times in order to study variation in the way people answered questions about choices under uncertainty.

Putting our research in context to the aforementioned literature, we are particularly interested in the following two questions. First, if context matters, so should the order of the questionnaires. Second, if happiness is to be a stable evaluation of one's life-as-a-whole, some deliberate, cognitive reasoning must be part of the response process. While we cannot directly observe the amount of cognitive activity, it is reasonable to take both response time and the "made me think" evaluation as proxies for it. This relates to the distinction between fast, intuitive System 1 responses and slow, systematic System 2 responses (e.g. Kahneman, 2011). Again, we want to find out whether this distinction has a discernible effect on reported happiness.

Any such effect, if present, does not necessarily constitute a problem for happiness research. As long as it can be treated as random, such as the weather or other determinants of transient mood, it just adds to the noisiness of the happiness measure, and the noise will tend to cancel out when averages are taken. Of course, the "randomness" assumption no longer holds, if the response context was purposefully manipulated to obtain certain answers. But even in absence of such a manipulation, one needs to be concerned if response context affects the sensitivity of reported happiness to socioeconomic characteristics (as shown by Strack et al., 1988). It could then well be the case that the relationship between reported happiness and socioeconomic characteristics is only a poor indicator for the relationship between underlying long-term happiness and those same characteristics.

The remainder of the paper is structured as follows: data collection, sample and variables are described in Section 3.2. In Section 3.3, the concept of a happiness reporting function is introduced. We specify a non-linear regression model and discuss estimation by quasi-maximum likelihood. Our results in Section 3.4 suggest that reporting behaviors affect levels of reported happiness and the perception of different happiness determinants. Importantly, response correlates moderate the trade-off ratio between income and unemployment. We discuss one particular theoretical framework that can explain our results. Section 3.5 concludes with a discussion of implications for ongoing happiness research.

## 3.2 Data

### 3.2.1 Happiness questionnaire

We use data from a monthly internet panel, the Longitudinal Study for Social Sciences (LISS). The LISS panel is a general purpose household survey that collects comprehensive information on income, employment, education, social participation and attitudes, among others. The panel survey is run by CentERdata, based at Tilburg University in the Netherlands. It was started in 2007, when 10'150 addresses were randomly drawn from the Dutch population register and 5176 households initially agreed to participate in the survey (see Scherpenzeel, 2009, for further details). Following a refreshment sample stratified by age, ethnicity and household types in 2009 the LISS panel has been shown to be representative for the Dutch population (de Vos, 2010). To the best of our knowledge, no comparable large-scale representative internet-based household panel survey exists anywhere else in the world.

The analysis of this paper is based on a happiness module that was in the field during March and April of 2011. We have 4399 valid responses of individuals participating either in the March or April wave. The happiness questionnaire consisted of four consecutive screens. Figure 3.1 shows the screenshots. The first page tells participants that only one question will be asked. The second page displays a usual single item happiness question. Participants answered the question "All things considered, how happy would you say you are?" on a Likert scale ranging from 0 to 9. Next, on the third page of the questionnaire, respondents were invited to evaluate the happiness question by assessing difficulty in answering, clarity of the question, and degree of thought provocation. On the last page participants were offered the possibility to write a comment. Only 35 individuals did so.

Furthermore, we have information on a participant's age, gender, income, employment and marital status, education, household composition and country of origin. Means and standard deviations of reported happiness and background variables employed in this study

are shown in Table 3.1. The average happiness is 7.15 on the 0-9 scale, and thus quite high. The employment rate is 52 percent, and as the average age of 51 indicates, the sample contains a sizeable proportion of retired individuals.

### 3.2.2 Reporting correlates

At the beginning of each month the LISS participants receive an electronic message including web links directing to different question modules. Participants can freely choose at which day and time or in which order they want to respond to the question modules. The LISS mechanically collects time stamp data on the interaction between the user and the underlying database. Thus, it is known for example, at what time a particular question module or question was opened and when an answer was sent back. Moreover, as stated above, participants self-assess their reporting behavior on page 3 of the happiness questionnaire. We use these data to construct three reporting correlates.

The first reporting correlate measures the time used to answer the happiness question. Figure 3.2 shows a kernel estimate of response times in seconds, using an Epanechnikov kernel with bandwidth equal to 0.6. Response times vary substantially, although 50% of all individuals answered within 8 seconds. The minimum answer time was 2 seconds and the maximum 97 seconds. The average answer speed is 10 seconds, which is similar to that reported by Couper et al. (2006) for other rating questions. It is unlikely that variation in response times is related to varying speed of the internet connection, as the LISS panel offers broadband internet connections to all participating households (Scherpenzeel, 2009).

A possible explanation for variation in response times is reading speed. Reading speed may also correlate with (observed and unobserved) determinants of happiness. We obtained adjusted response times by estimating an exponential regression model of item response times. The estimated marginal effects of a large set of socioeconomic and sociodemographic variables are shown in Table 3.2. Older people and foreigners tend to answer more slowly, whereas employed, married and better educated participants respond faster on average. The

response time is also higher for those who mentioned difficulties in answering. The residual of this first stage regression gives us the adjusted response time for each individual. This is our first and main reporting correlate. It is positive if a person was slower in answering the happiness question than a typical person with similar characteristics, and negative otherwise. One possible interpretation of this adjusted answering time is that it proxies for the amount of thinking, or deliberate cognitive effort, individuals put into answering the question.

The questionnaire offers a second potential piece of information regarding cognitive effort, namely self-assessed responses of people how they went about answering the happiness question. In particular, page three of the questionnaire included the question: "Did the (happiness) question get you thinking about things?". The answer scale went from 1 (="totally disagree") to 5 (="totally agree"). More than 20% of the respondents completely disagreed with this statement, and about 12% disagreed. We construct a dichotomized version of this variable, "cognitive effort", taking the value 0 if a respondent either disagreed or strongly disagreed, and the value 1 else. Of course, this is not an objective measure of cognitive effort but rather subjective and self-assessed. Also, by its very nature it is limited to the conscious dimension of effort. This question has been included by LISS originally for the purpose of questionnaire development. We are not aware of any prior research using this information in the context of happiness equations.

The third reporting correlate captures the context of the answering process. The LISS panel sends each month a Background Variable Questionnaire (BVQ) to the contact person of the household. We compared the activation time of the BVQ to that of the happiness questionnaire. The third reporting correlate takes the value 1 if the BVQ was opened by the contact person during the two hours preceding the happiness questionnaire. It is zero otherwise. Ones are observed in about 23% of all cases. There are a number of possible hypothesis why there might be an effect on reported happiness. One is salience of the information that was provided in the BVQ. The other is the effect on response burden

which markedly increases for those who chose to answer the BVQ first. For example, Galesic and Bosnjak (2009) have shown that respondents change their response behavior (faster responses and less variation) with increasing time spent on a questionnaire.

## 3.3 Models

### 3.3.1 Reporting function

Suppose that "true" happiness $H$, i.e. a person's evaluation of life-as-a-whole absent of distorting transient influences, depends on a vector of external factors $x_1, \ldots, x_k$ such that $H = h(x_1, \ldots, x_k)$. Moreover, let reported happiness $R$ be given by

$$R = r(x_1, \ldots, x_k, z_1, \ldots, z_m) + \varepsilon \tag{3.1}$$

where $z_1, \ldots, z_m$ denote reporting correlates and $\varepsilon$ captures the influence of transient influences, assumed to be independent of reporting correlates and happiness determinants. In contrast to Oswald (2008) who studied the shape of the reporting function $r$, the focus here is on possible interactions between $x_j$ and $z_l$. In general, it will be the case, because of such interactions, that $\partial H/\partial x_j \neq \partial R/\partial x_j$, and the marginal rates of substitution, and trade-off ratios, differ for $H$ and $R$. However, suppose instead that (3.1) can be re-written as

$$R = r(h(x_1, \ldots, x_k), z_1, \ldots, z_m) + \varepsilon \tag{3.2}$$

This is a special case of (3.1), where $x_j$ affects reported happiness only through $h(x_1, \ldots, x_k)$. Model (3.2) implies that $z$ does not moderate the effect of $x_j$ on $R$, since we can write the the marginal rate of substitution, or relative marginal effects, as

$$\frac{\partial R/\partial x_i}{\partial R/\partial x_j} = \frac{\partial R/\partial H \cdot \partial H/\partial x_i}{\partial R/\partial H \cdot \partial H/\partial x_j} = \frac{\partial H/\partial x_i}{\partial H/\partial x_j} \tag{3.3}$$

Under model (3.2), reported happiness identifies the relative marginal effect of true happiness. Under model (3.1), this is not the case, as the reporting process drives a wedge

between true and reported happiness that distorts relative effects. A key objective of this paper is therefore to test model (3.2) against the more general model (3.1). If the reporting function (3.2) cannot be rejected, then we know that the difference between reported and true happiness is unimportant, as long as conclusions focus on relative effects of true happiness.

### 3.3.2 Empirical model

The deterministic part of a linearized version of the reporting model can be written as

$$R = x'\beta + z'\gamma + (zx)'\delta \tag{3.4}$$

where $\delta$ captures interaction effects. The effect of socioeconomic background variables on reported happiness is a function of reporting correlates as long as $\delta \neq 0$. For instance, with $z$ being the adjusted response time and $x$ income, $\beta > 0$ and $\delta < 0$ would imply that the marginal effect of income on reported happiness is higher when the answer is given more slowly. In other words, respondents would attribute the less weight to income the longer they take to answer. Hence, it depends on $\delta$ whether the reporting function has the form of model (3.1) or model (3.2). If $\delta$ is a multiple of $\beta$, relative marginal effects with respect to components of $x$ are unchanged by $z$, giving rise to model (3.2). Otherwise, model (3.1) is obtained.

Since $R$ is logically restricted to lie between 0 and 9, it is impossible to observe negative mean values $E(R|x, z)$, or values above 9. This consideration would be ignored by a linear regression model. Hence we specify a non-linear regression model whereby

$$R = f[x'\beta + z'\gamma + (zx)'\delta] + v\,, \tag{3.5}$$

$f$ is a transformation function that maps the real line onto the $[0, 9]$ interval, and $E(v|x, z) = 0$. This model is of the form of a Rating Scale Model, which is detailed in Section 1.3.3 of this dissertation. We use Bernoulli quasi-maximum likelihood estimation with the transformation function specified as a modified logit function. Marginal effects differ from individual

to individual due to the non-linearity. As a rule of thumb, average marginal effects can be obtained by multiplying the coefficient with the factor $\bar{R}(R^{max} - \bar{R})/R^{max}$, where $\bar{R}$ is mean reported happiness in the sample. In our data, this factor is approximately equal to 1.5.

## 3.4 Results

We first use our data to estimate a happiness equation without including reporting correlates. $x$ includes socioeconomic individual determinants of happiness that are commonly used in the economic well-being literature (e.g., Frey and Stutzer, 2002). Estimates of the parameter vector shown in column 1 of Table 3.3 replicate standard findings. Reported happiness is positively associated with income, marriage and employment. Men and foreigners report lower happiness and happiness scores are U-shaped in age. The magnitude of the associations are similar to earlier findings as well. For instance, a 1% raise in income is associated with an increase of reported happiness by 0.4 points on average.

Columns 2 to 4 in Table 3.3 add one reporting correlate at a time to the regression, assuming the absence of interaction terms (i.e., $\delta = 0$). Table 3.3 provides evidence that reporting correlates do correlate with levels of reported happiness, ceteris paribus. For instance, based on column 2, an increase in adjusted response time by 12 seconds is associated with an approximately 0.1 point lower reported happiness. A difference of similar magnitude results when comparing respondents with a cognitive effortful answer and those without. These associations might look small at first. However, if compared to effects of other socioeconomic characteristics, they are actually quite large. For instance, they are larger in absolute value than the impact of being employed versus non-employed (this includes unemployment and non-participation). The last column of Table 3.3 reports a positive association of questionnaire order with reported happiness. Those, who reviewed the background questionnaires first report a higher happiness, on average. Another note-

worthy feature of the results in Table 3.3 is that the estimated parameter vector for the socioeconomic determinants is relatively insensitive to the inclusion of reporting correlates. Hence, a regression of reported happiness on individual characteristics excluding reporting correlates seems not suffer from omitted variable bias.

Results shown in Table 3.4 refer to the unconstrained empirical model (3.5). The estimates tell us whether reporting circumstances change the estimated relationship between happiness and these socioeconomic characteristics, and in particular, whether relative marginal effects change (i.e., the distinction between reporting function (3.1) and (3.2). The upper panel of the table reports estimates of the main effects of happiness determinants ($\hat{\beta}$). The lower part of the table displays the main effect of the reporting correlates together with the estimated interaction coefficients $\hat{\delta}$. For the sake of exposition, Table 3.4 includes only part of the socioeconomic coefficients and interaction parameters, although the models were estimated with the same set of variables that were used in Table 3.3 (not shown are the main effects and interactions of age, age$^2$, male, foreigner, log number of household members, and April interview).

Again, the analysis is done separately for the three correlates. Column 1 of Table 3.4 shows the results for the happiness equation that is interacted with response time. All but one of the interaction terms are close to zero and statistically insignificant. The exception is the effect of income on reported happiness that is found to decrease with response time. For instance, the average marginal effect of a 1% income increase increases by 0.04 points, or 15%, if the response time is reduced by 10 seconds. An even stronger interaction effect of income is found in column 2 of Table 3.4, where the marginal effect of a 1% income increase is more than twice as large for those individuals who stated that answering the question required no cognitive effort, as opposed to others.

The last column of Table 3.4 shows results for the questionnaire order variable. It is conceivable that answering the socioeconomic questions increases the salience of these variables, leading to a stronger observed relationship. Also, questionnaire order might lead

to priming (Strack at al., 1988), whereby participants substitute answers given to previous questions, for instance about their income or employment status, for the assessment of happiness. However, we cannot find any evidence for such an effect in our data.

Summarizing the evidence, we find a statistically significant interaction effect in the model (3.2) regressions, but only for income and only for response time and cognitive effort. Specifically, slower and more thoughtful answers reduce the happiness-income gradient. Since the evidence suggests therefore a non-proportional moderation of the income effect relative to the effects of other socioeconomic characteristics on happiness, marginal rates of substitution, or trade-off ratios, might not be invariant to response behavior.

## A possible explanation

Suppose, as in Rubinstein (2007), that there are two polar states of mind for answering happiness questions, an instinctive one and a cognitive one. These polar states are also referred to as System 1 and System 2 processes in cognitive psychology and decision theory (e.g. Stanovich and West, 2000, Frederick, 2005). As a general happiness question is of an evaluative nature, one might expect that those who think longer about their answer and also state that they spent more cognitive effort, are the same individuals, for which the happiness answers are less random and for whom one finds stronger relationships to the socioeconomic determinants. However, this is not what we find, on the contrary.

A possible resolution to this "puzzle" is an alternative conception, whereby reporting correlates, and the response time in particular, do not primarily relate to the amount of cognitive deliberations when answering the question, but rather proxy for mood (e.g., Frederick, 2005; Kahneman, 2011). People in a good mood are more likely to answer spontaneously and intuitively, while people in a bad mood are more likely to rely on effortful mental activities when answering a question.

Such an explanation would be compatible with our observation that slower respondents, as well as those exerting more cognitive effort, report lower levels of happiness. In this

interpretation, the reduced happiness does not result from the higher effort per se, but rather because high effort proxies for bad mood which is otherwise not captured by the model. It also appears that a negative mood reduces the weight that individuals give to income changes when thinking about their happiness. The reason for this phenomenon is less clear and remains an interesting question for future research.

## 3.5 Conclusions

When asked to respond to a survey question on how happy a person is with her life, some people respond quickly and some take more time to respond. Some people say that the happiness question got them thinking, while others don't. The objective of this paper was to explore whether these circumstances of reporting are associated with reported happiness, using data from a Dutch internet panel. There were two main findings on response time. First, responding slowly is associated with a lower reported happiness. A possible explanation is that respondents in a positive mood are more likely to give intuitive and therefore faster, answers.

Second, the marginal effect of income on happiness decreases with response time. To illustrate the magnitude of this effect, we can compute the income compensation that is necessary in order to make a non-working person equally well off to an employed person in terms of reported happiness. For a person with an average response speed this estimated compensation amounts to 32% of the initial income. For a person, who takes a standard deviation longer to answer, the estimated compensating income is 37% of the initial income.

Happiness research finds itself at a critical juncture, where results are increasingly used to inform and formulate policy interventions. A key promise of happiness research for such policy debates is that it allows to overcome the limitation of traditional cost-benefit analysis that everything has to be measured up in dollars. With the increasing availability of happiness data, it becomes possible, at least in principle, to value policy trade-offs in

terms of their effect on happiness, well-being or utility. Our paper suggests a possible limitation to this approach, as our findings highlight the possibility that such measured trade-off ratios may not be invariant to the circumstances of reporting.

# References


Couper, M. P., R. Tourangeau, F. G. Conrad and E. Singer, 2006, "Evaluating the Effectiveness of Visual Analog Scales : A Web Experiment", *Social Science Computer Review*, 24, 227-245

de Vos, K., 2010, "Representativeness of the LISS-panel 2008, 2009, 2010", published online http://www.lissdata.nl, last consultation 14.10.2011

Diener, E., 1984, "Subjective Well-Being", *Psychological Bulletin*, 95, 542-575

Frederick S., 2005, "Cognitive Reflection and Decision Making", *Journal of Economic Perspectives*, 19, 4, 25-42

Frey, B. S. and A. Stutzer, 2002, "Happiness and Economics: How the Economy and Institutions Affect Human Well-Being", Princeton: Princeton University Press

Galesic, M. and M. Bosnjak, 2009, "Effects of questionnaire length on participation and indicators of response quality in a web survey", *Public Opinion Quarterly*, 73, 349-360

Kahneman, D., 2011, "Thinking, fast and slow", New York: Farrar, Straus and Giroux

Krueger, B. and D. A. Schkade, 2008, "The Reliability of Subjective Well-Being Measures", *Journal of Public Economics*, 92, 1833-1845

Oswald, A., 2008, "On the curvature of the happiness reporting function from objective reality to subjective feelings", *Economic Letters*, 100, 3, 369-372

Pavot, W., and E. Diener, 1993, "The affective and cognitive context of self-reported measures of subjective well-being", *Social Indicators Research*, 28, 1-20

Rubinstein, A., 2007, "Instinctive and Cognitive Reasoning: A Study of Response Times", *The Economic Journal*, 11, 1243- 1259

Scherpenzeel, A., 2009, "Start of the LISS panel: Sample and recruitment of a probability-based Internet panel", published online http://www.lissdata.nl, last consultation 14.10.2011

Schwarz, N., 1987, "Stimmung als Information: Untersuchungen zum Einfluss von Stimmungen auf die Bewertung des eigenen Lebens", Heidelberg: Springer Verlag

Schwarz, N. and G. L. Clore, 1983, "Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States", *Journal of Personality and Social Psychology*, 45, 3, 513-523

Stanovich, K. E. and R. F. West, 2000, "Individual Differences in Reasoning: Implications for the Rationality Debate?", *Behavioral and Brain Sciences*, 22, 5, 645-726

Strack, F., L. L. Martin and N. Schwarz, 1988, "Priming and communication: Social determinants of information use in judgments of life satisfaction", *European Journal of Social Psychology*, 18, 429-442

Veenhoven, R., 1984, *Conditions of Happiness*, Dordrecht: D. Reidel

Table 3.1: Descriptive statistics

| | Mean | Standard deviation |
|---|---|---|
| Happiness | 7.15 | 1.19 |
| Log after-tax household income | 7.84 | 0.52 |
| Proportion employed | 0.52 | 0.50 |
| Proportion with higher education | 0.53 | 0.50 |
| Age | 51.00 | 16.97 |
| Proportion male | 0.47 | 0.50 |
| Proportion married | 0.59 | 0.49 |
| Proportion foreigners | 0.12 | 0.33 |
| Log HH members | 0.80 | 0.52 |
| Proportion interviewed in April | 0.48 | 0.50 |

$N = 4399$, Source: Longitudinal Study for Social Sciences, 2011

Table 3.2: Exponential regression of response time on characteristics

| | Marginal effect | Standard error |
|---|---|---|
| Male | 0.183 | (0.208) |
| Age | 0.098*** | (0.008) |
| Log after-tax household income | -0.485 | (0.455) |
| Log number of household members | 0.094 | (0.379) |
| Houseownership | -0.493 | (0.300) |
| Employed | -1.212*** | (0.275) |
| Secondary education | -0.825** | (0.365) |
| Tertiary education | -0.979** | (0.381) |
| Married | -0.626** | (0.305) |
| Cohabiting | 0.016 | (0.483) |
| Separated | -0.327 | (0.461) |
| Foreigner | 1.535*** | (0.423) |
| Returned to the question | 3.126* | (1.795) |
| Difficulty in answering | 0.455*** | (0.103) |
| April interview | 0.141 | (0.207) |

$N = 4399$; Heteroscedasticity consistent standard errors presented in parentheses. ***, **, * denote statistical significance at the 1%, 5%, 10% significance levels, respectively.

Table 3.3: Regressions of reported happiness on characteristics and reporting correlates

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Response time | | -0.007*** | | |
| | | (0.002) | | |
| Self-assessed cognitive effort | | | -0.083*** | |
| | | | (0.026) | |
| Background questionnaire first | | | | 0.056* |
| | | | | (0.031) |
| Log after-tax household income | 0.265*** | 0.265*** | 0.264*** | 0.267*** |
| | (0.030) | (0.030) | (0.030) | (0.030) |
| Employed | 0.078** | 0.078** | 0.078** | 0.077** |
| | (0.032) | (0.032) | (0.032) | (0.032) |
| Tertiary degree | 0.006 | 0.004 | 0.007 | 0.002 |
| | (0.026) | (0.026) | (0.026) | (0.026) |
| Age | -0.028*** | -0.028*** | -0.028*** | -0.029*** |
| | (0.005) | (0.005) | (0.005) | (0.005) |
| Age$^2$ $\times 10^{-2}$ | 0.030*** | 0.031*** | 0.030*** | 0.031*** |
| | (0.005) | (0.005) | (0.005) | (0.005) |
| Male | -0.070*** | -0.068*** | -0.071*** | -0.068*** |
| | (0.024) | (0.024) | (0.024) | (0.024) |
| Married | 0.317*** | 0.318*** | 0.314*** | 0.320*** |
| | (0.032) | (0.032) | (0.032) | (0.032) |
| Foreigner | -0.168*** | -0.167*** | -0.165*** | -0.170*** |
| | (0.038) | (0.038) | (0.038) | (0.038) |
| Log number household members | -0.100*** | -0.100*** | -0.096*** | -0.096*** |
| | (0.032) | (0.032) | (0.032) | (0.032) |
| April interview | -0.029 | -0.029 | -0.023 | -0.011 |
| | (0.024) | (0.024) | (0.024) | (0.026) |
| Constant | -0.230 | -0.227 | -0.176 | -0.254 |
| | (0.252) | (0.248) | (0.254) | (0.251) |

$N = 4399$; Estimates for the parameter vectors $\beta$ and $\gamma$ of model (3.5) are shown. Heteroscedasticity consistent standard errors presented in parentheses. ***, **, * denote statistical significance at the 1%, 5%, 10% significance levels, respectively.

Table 3.4: Reported happiness, socioeconomic characteristics, and reporting correlates

| | (1) | (2) | (3) |
|---|---|---|---|
| Log household income | 0.274*** | 0.423*** | 0.249*** |
| | (0.030) | (0.057) | (0.034) |
| Employed | 0.077** | 0.020 | 0.074** |
| | (0.032) | (0.061) | (0.036) |
| Tertiary degree | 0.005 | -0.040 | 0.015 |
| | (0.026) | (0.049) | (0.030) |
| Married | 0.319*** | 0.265*** | 0.297*** |
| | (0.032) | (0.060) | (0.037) |
| Response time | 0.046* | | |
| | (0.024) | | |
| Log household income | -0.004* | | |
| × Response time | (0.002) | | |
| Employed | 0.000 | | |
| × Response time | (0.004) | | |
| Tertiary degree | 0.001 | | |
| × Response time | (0.004) | | |
| Married | 0.002 | | |
| × Response time | (0.005) | | |
| Cognitive effort | | 1.912*** | |
| | | (0.544) | |
| Log household income | | -0.222*** | |
| × cognitive effort | | (0.066) | |
| Employed | | 0.079 | |
| × cognitive effort | | (0.072) | |
| Tertiary degree | | 0.070 | |
| × cognitive effort | | (0.058) | |
| Married | | 0.075 | |
| × cognitive effort | | (0.071) | |
| Background questionnaire first | | | -0.538 |
| | | | (0.598) |
| Log household income | | | 0.086 |
| × Background questionnaire first | | | (0.073) |
| Employed | | | 0.031 |
| × Background questionnaire first | | | (0.083) |
| Tertiary degree | | | -0.059 |
| × Background questionnaire first | | | (0.063) |
| Married | | | 0.113 |
| × Background questionnaire first | | | (0.075) |

$N = 4399$; Estimates for the parameter vectors $\beta$, $\gamma$ and $\delta$ of model (3.5) are shown. Heteroscedasticity consistent standard errors presented in parentheses. ***, **, * denote statistical significance at the 1%, 5%, 10% significance levels, respectively. The models include in addition the variables age, age$^2$, male, foreigner, log household members, April interwiew and their interactions with the respective reporting correlates.

Figure 3.1: Screenshots of happiness questionnaire

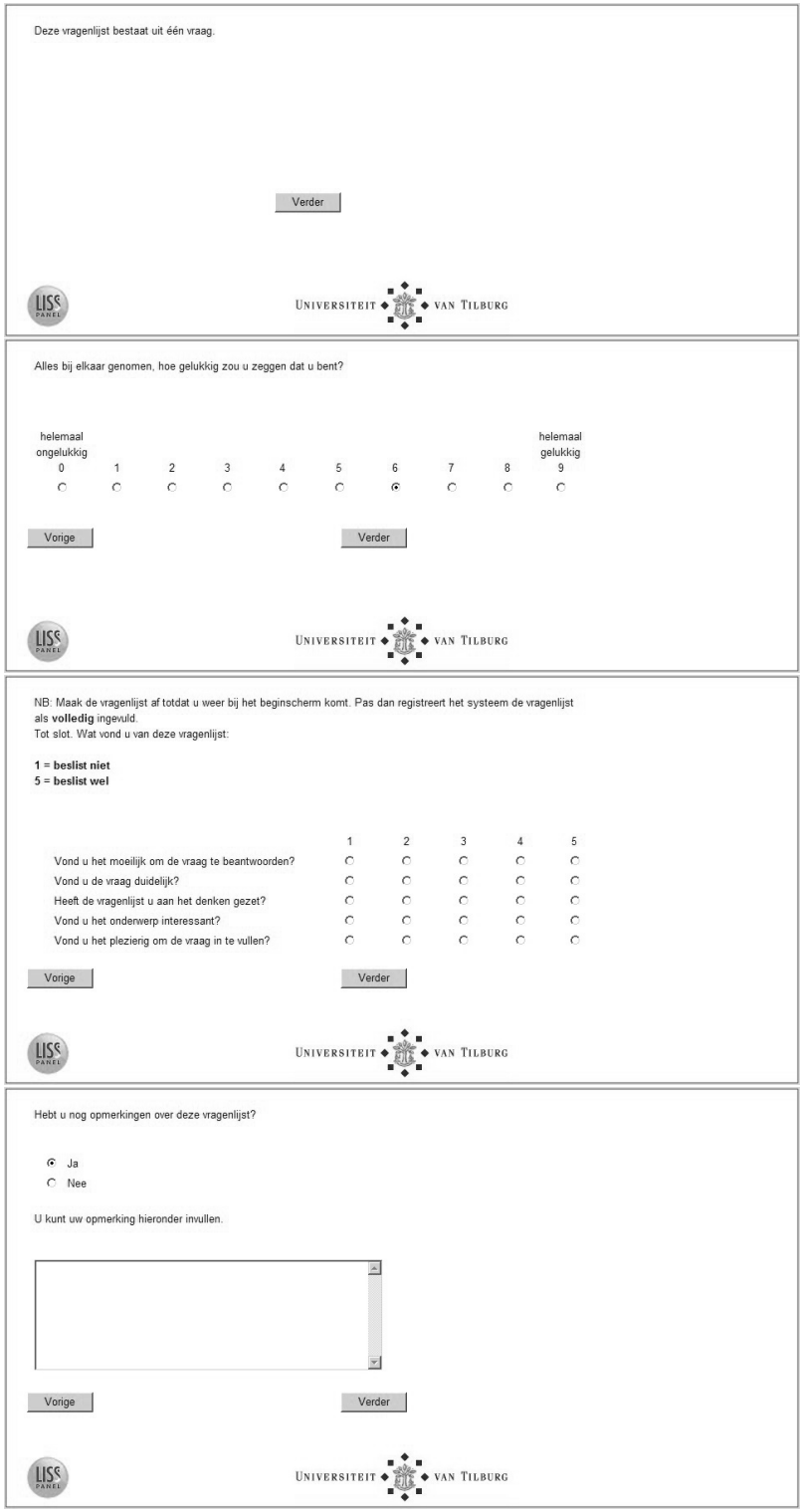
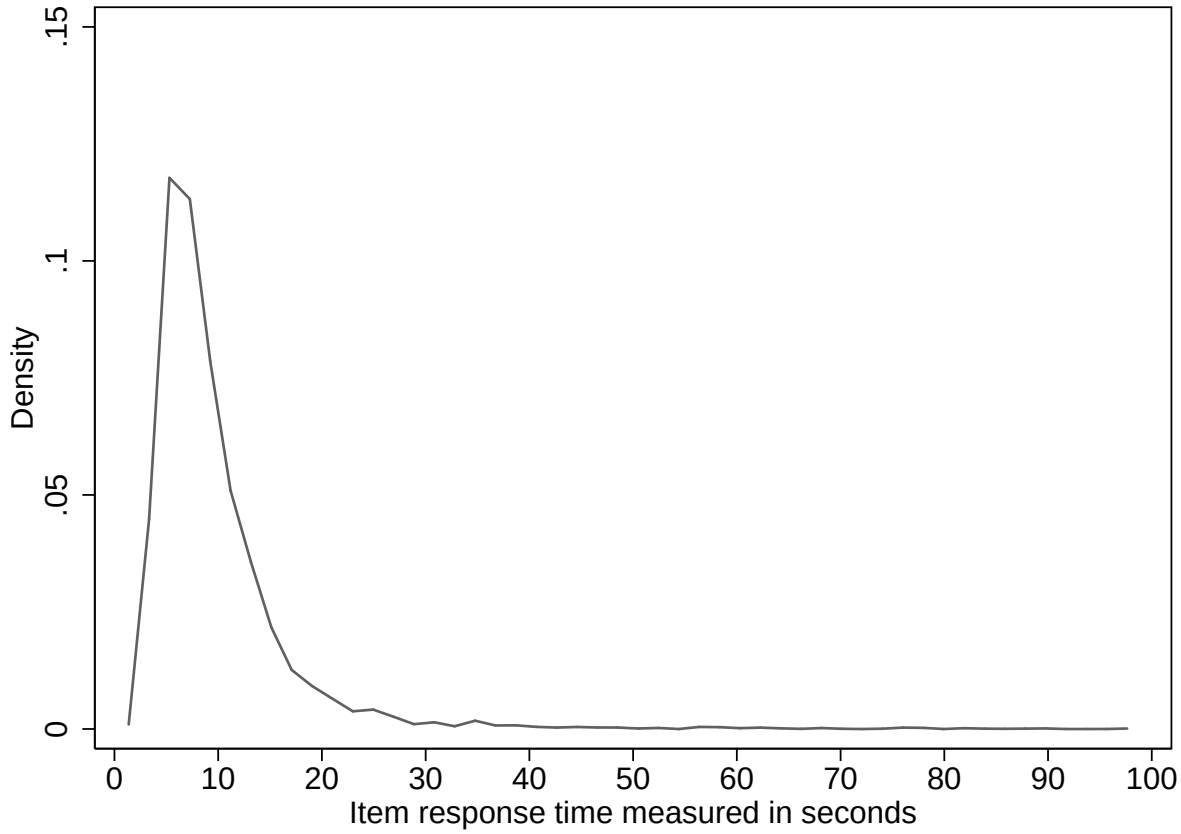


Figure 3.2: Kernel density estimate of response time invested in happiness question




*Notes:* $N = 4399$; An Epanechnikow kernel with bandwidth equal to 0.6 is used to estimate densities.

# Chapter 4

# Does the stork deliver happiness? Parenthood and life satisfaction

This chapter is joint work with Gregori Baetschmann and Kevin E. Staub. It is similar to Working Paper No. 94 published in the *Working Paper Series* of the Department of Economics, University of Zurich.

*Acknowledgements*: We thank Timo Boppart, Arie Kapteyn, Giovanni Mastrobuoni, Andreas Steinhauer, Rainer Winkelmann and participants at the Zurich Workshop for Economics and the RAND Labor & Population Brown Bag Seminar for comments. Kevin E. Staub acknowledges financial support from the Swiss National Science Foundation through grant PBZHP1-138692.

## 4.1 Introduction

How does becoming a mother affect women's life cycle utility streams? Rational choice approaches to fertility embedded in standard dynamic economic models of fertility assume that the net utility gain of motherhood is positive. In sharp contrast, the predominant view in the sociological and psychological literature is that there is a negative net effect of parenthood. This view is derived from the empirical literature on subjective well-being where the correlation between having children and life satisfaction is usually found to be negative.

In the previous literature, the implicit control group for parents is represented by childless individuals with the same covariates. This empirical strategy is problematic if parents differ from non-parents in terms of unobserved qualities. One preeminent possible source for such differences in the context of parenthood is self-selection. We show that selection on observable and unobservable characteristics into parenthood is indeed important: prospective mothers' satisfaction increases around five years before first delivery. This suggests the use of exogenous variation in fertility choices to estimate the gains in life satisfaction derived from becoming a parent. Exogenous variations which have been shown to impact fertility decisions include job displacements (Del Bono et al., 2012) or the homogeneity of the first two children's sex (Angrist and Evans, 1998), for instance. However, while such variation is unlikely to be correlated with a number of outcomes of interest, it seems difficult to argue that it does not affect mothers' life satisfaction. Thus, to answer the question of how individual well-being is affected by parenthood alternative empirical strategies need to be explored. The key contribution of this paper is to propose regression models which —exploiting either intra- or interpersonal variation– embed differences in unobserved characteristics that are likely to increase the likelihood of motherhood. In our preferred specification, for instance, we match prospective mothers to women who will never have children but who are similar to prospective mothers in terms of past life satisfaction paths and observable characteristics. Our results suggest that motherhood is associated with a

substantial net utility gain, a finding consistent with rational choice approaches to fertility.

Broadly, this paper contributes to the strand of the literature on the economics of happiness which aims at providing (rough) estimates of trade-offs guiding choice behavior.[1] The last decade has seen a boom in the field of happiness economics with a diverse host of both theoretical and empirical contributions.[2] One reason for this growth has been the increasing evidence from economists and psychologists alike suggesting that individual responses on subjective well-being collected from surveys can be usefully interpreted as proxy measures for utility in a variety of contexts.[3] While the issue studied most intensely has been the relationship of income and employment to well-being, other aspects such as health, marriage and religion have also received due attention in the literature. In each of these cases, the existing research has been able to uncover clear satisfaction gains associated with these factors as would be expected from a mainstream view of utility.[4]

Fertility, by contrast, is an aspect which has received less direct attention in the happiness literature, at least relative to its important place in microeconomic theory and extensive body of accompanying empirical research dating back to Becker (1960) and Willis (1973). The predominant finding across numerous datasets is that individuals with children report on average lower satisfaction than comparable childless adults. This negative correlation has found ample resonance in some strands of the sociological and psychological literature, where the result is usually interpreted as a negative net effect of parenthood. Two main rationalizations have been put forward to explain why most adults select into parenthood despite costs apparently outweighing benefits. The first explanation, com-

[1] Following the convention in economics, we use the words happiness, satisfaction and well-being as synonyms.

[2] See Ferrer-i-Carbonell (2012), Blanchflower (2009), Layard (2005), Frey and Stutzer (2002) and Kahneman, Diener and Schwarz (1999) for surveys of this literature.

[3] An in-depth review on the literature linking subjective well-being to utility can be found in Clark, Frijters and Shields (2008). See Benjamin et al. (2012) for a recent contribution.

[4] The seminal paper in the literature on income and happiness is Easterlin (1973); see Easterlin (2001) and Stevenson and Wolfers (2008) for recent additions. For sources on the literature on unemployment we refer to Clark and Oswald (1994) and Winkelmann and Winkelmann (1997). For contributions on the relationship between happiness and marriage, and happiness and health, see e.g. Stutzer and Frey (2006) and Veenhoven (2008), respectively.

mon in the sociological literature, emphasizes the presence of pro-natal social norms which sanction disconformity (Morgan and Berkowitz King, 2001; Vanassche, Swicegood and Matthijs, 2012). The second, psychological explanation sees the choice for having children as an instance of biased affective forecasting, i.e. individuals making rational decisions based on incorrect expectations (Gilbert, 2006) – in this case, based on the widespread belief expressed in surveys that having children brings happiness (Hansen, 2011). Among economists, on the other hand, the finding has been treated with more reservation, and few attempts at rationalizing it have been undertaken.[5] However, the negative correlation is acknowledged regularly in survey articles in the economic literature (Blanchflower, 2009; Clark, Frijters and Shields, 2008; Dolan, Peasgood and White, 2008; Ferrer-i-Carbonell, 2012), and incidental interpretations along the lines of the psychological and sociological research are not uncommon.

Much of what is known on the subject does not stem from studies focusing on fertility; rather it often comes from regression studies where fertility measures are used as controlling variables to avoid confounding a specific effect of interest (Di Tella, MacCulloch and Oswald, 2001, 2003; Alesina, Di Tella and MacCulloch, 2004; Clark, 2007). Three frameworks have been used to study the effect of parenthood on life satisfaction: (i) cross-section and pooled panel regression models, (ii) panel models with fixed effects and (iii) event studies. By far the most common of these is the first framework. Recently, Stanca (2012) confirmed the presence of the negative parenthood effect using this standard happiness equation framework for over 90 countries. Herbst and Ifcher (2012) closely scrutinize the negative effect obtained with this framework for US data, concluding that the magnitude of the effect has been decreasing in the last decades and that it is driven mainly by older parents. The negative effect has also been found using the second framework (e.g. Stutzer and Frey, 2006). In the few instances where the association is found to be positive, it is usually

[5] The small strand of the economic happiness literature focusing on life event studies is an exception in this respect (Clark et al., 2008, Frijters, Johnston and Shields, 2011). These papers, too, find little evidence for a parenthood effect, but they explain their result with adaptation, a concept derived from set point theory. We discuss these findings in more detail below.

small and insignificant (Clark and Oswald, 2002).[6] The third approach is life-event studies tracking parental satisfaction over a time window around the birth of a child (Clark et al., 2008, Frijters, Johnston and Shields, 2011). This research has concluded that parents adapt completely to the birth of a child after a brief time; i.e. heightened happiness levels return to a previous baseline level, sometimes even dipping below the baseline.

The estimation approaches (i), (ii) and (iii) used by the previous literature are inadequate to measure utility gains from parenthood. A first concern relates to the insight from standard dynamic economic models of fertility which suggest that other outcome variables such as income, partnership status and employment are endogenous to the fertility decision.[7] An implication hereof is that the ceteris-paribus effects reported in the previous literature are difficult to interpret. These effects represent an ex-post comparison of satisfaction between parents and individuals with no children at the same values of other outcomes, when optimally these outcomes will differ precisely as a consequence of the parenthood decision.[8] Indeed, Herbst and Ifcher (2012), who extensively assess the robustness of the traditional happiness-equation estimates of the parenthood effect, find that the estimates are quite sensitive to the inclusion of different sets of covariates, a typical result when conditioning on mediator variables which are part of the channels through which the effect runs.

The second important concern relates to the selection into motherhood. In approach (i), most of the individuals observed without children are on their way of becoming parents. The self-selection we identify in our analysis implies that using such prospective parents' satisfaction as a counterfactual outcome for parenthood is misleading. In this standard approach, prospective parents are censored and their outcomes attributed to non-parents, and

[6] One of the few studies reporting a significant positive association is Kohler, Behrman and Skytthe (2005) who study identical twins.

[7] Arroyo and Zhang (1997) provide an overview of the early dynamic fertility model literature; for an example of contemporary research encompassing occupational choice, marriage and fertility, see Ma (2010). Recent studies focusing explicitly on motherhood are surveyed in Del Boca and Locatelli (2006), see also Wilde, Batchelder and Ellwood (2010) and Michaud and Tatsiramos (2011).

[8] Figure C.1 in the Appendix illustrates this point by plotting working hours over the life cycle for women remaining childless and mothers with age at first birth 28.

therefore the average satisfaction level of childless adults is overestimated. Moreover, the dynamics of self-selection we find also affect approaches (ii) and (iii). In these approaches the effect of parenthood is identified by comparing pre- and post-birth satisfaction levels of mothers. Given the heightened pre-birth happiness of mothers during the five years foregoing first birth, individual fixed effects are biased upwards and induce a negative bias in the effect of interest. In life event studies this distortion is amplified because such studies usually use a window of only two or four years around the event "birth of a child."

A careful study into the effect of motherhood on satisfaction needs to account for these methodological issues, and we propose estimation strategies which do so. First, we construct a completed fertility decision sample consisting of women whose completed fertility is observed. This ensures the correct classification of women which are about to become mothers (to whom we simply refer to as mothers henceforth) and of women which are never to have children (to whom we refer to as non-mothers). Second, we establish comparability on observable characteristics, such as income, partnership status, etc., between mothers and non-mothers *before* mothers first gave birth to a child. Third, and most important, we account for the five-year-long increase in mothers' life satisfaction that precedes birth of the first child with two different identification strategies. On one hand, we construct a suitable control group for mothers from comparable non-mothers who experienced a satisfaction path similar to that of mothers before first birth. On the other hand, we compare mothers' life satisfaction after birth to their own life satisfaction levels before the onset of the five-year selection period.

For both these approaches we estimate the effect of motherhood for every year from first pregnancy to twenty years after transition to motherhood. We find the satisfaction gain of mothers to be positive throughout. The results are robust and similar for the various estimation strategies we propose, including nearest-neighbor matching and regressions with and without fixed effects, confirming the importance to account for self-selection into motherhood. Large effects occur in the first years after transition to motherhood and are

followed by a stabilization at a moderate level. We use the estimates to obtain a monetized net present worth of motherhood, finding the compensating variation of motherhood to lie roughly between one and two net yearly household incomes, depending on the estimates and discount rates used.

This chapter is organized as follows. In Section 4.2 we investigate selection into motherhood. Our methodological approaches tackling selection into motherhood are explained in Section 4.3. Section 4.4 contains our main regression results, and compares them to results obtained using traditional approaches. We explore further aspects related to fertility and life satisfaction as well, such as the effects at different ages of first birth, the effect for single-child and multiple-parity mothers, and the effect among fathers. Section 4.5 contains a concluding discussion.

## 4.2 Self-selection into motherhood

We use data on women from the German Socio-Economic Panel (GSOEP). The extended time dimension of the panel (twenty-five years in total) allows us to observe long periods of women's lives. In particular, we are able to identify women which later end up with a completed fertility of zero and study their satisfaction including the period of fertile years. The dashed line in Figure 4.1 presents the average satisfaction path of such non-mothers. Life satisfaction decreases until about the age of 55, and increases afterwards.[9] The solid line plots satisfaction of mothers delivering their first child at age 28. While satisfaction paths are similar after the age of 40, mothers' life satisfaction shows a pronounced peak around the year of first child's birth. Such an evolution of the satisfaction path is quite typical for mothers. The peak would be blurred, however, if the average satisfaction path for mothers with different ages at first birth was plotted.

Mothers' satisfaction path in Figure 4.1 is also clearly above non-mothers' path before

[9] Such U-shapes of satisfaction-age curves are common in the literature, cf. Van Landeghem (2012) and Wunder et al. (2011) for recent overviews.

and after transition into motherhood. While in this raw contrast the positive difference after first birth hints at possible satisfaction gains of motherhood, the pre-birth differences suggest that a more rigorous analysis of self-selection of mothers is needed.

We examine differences in pre-birth life satisfaction to study whether there is positive or negative selection on unobservable qualities conditional on observable characteristics. Again, we focus on women with observed completed fertility. Fertility is defined as completed by age 41. In our data, 99.8 percent of all mothers had given birth by that age. To identify the evolution before first birth precisely, we use information on the month of first child's birth and the months in which prospective mothers were surveyed in the years prior to first birth. This allows us to compute time to first birth in months. Details on the data are given in Appendix A.3.

We regress self-reported life satisfaction on indicators of number of months to first birth and control variables:

$$ls_{it} = \alpha + \mathbf{months\ to\ birth}'_{it}\beta + \mathbf{age}'_{it}\gamma + \mathbf{x}'_{it}\delta + \varepsilon_{it}, \tag{4.1}$$

where $ls_{it}$ is life satisfaction for individual $i$ in wave $t$ on a 11 points Likert scale. The vector $\mathbf{months\ to\ birth}_{it}$ consists of dummy variables, one for each month before first birth. An element takes the value one if a mother was surveyed during that specific month before birth of her first kid. All elements of $\mathbf{months\ to\ birth}_{it}$ are equal to 0 for non-mothers. The regression controls for age with a full set of dummy variables $\mathbf{age}_{it}$. Accounting flexibly for age is indispensable in the context of fertility. The vector $\mathbf{x}_{it}$ includes further control variables.[10] The variable $\varepsilon_{it}$ is the regression error.

Figure 4.2 visualizes the estimates of the parameters of interest in model (4.1) for the last seven years before first birth. The solid line shows average predicted life satisfaction for mothers. The dashed and dotted lines depict predicted life satisfaction for non-mothers using the covariate distribution of mothers. The regressions represented by the dashed and

[10] The further control variables are: survey year, number of years in panel, education, relationship status, household members, working hours and household income. Appendix A.1 contains a detailed description of the included terms.

dotted lines differ by the number of included control variables. Whereas the former only controls for survey year and years in panel, the regression of the dotted line also controls for the full set of socioeconomic controls. There is little difference between mothers' and non-mothers' life satisfaction until five years before birth. From that point on mothers' satisfaction increases steadily. The growth of the satisfaction path steepens around one year before birth. Women surveyed in the month before birth of their first child report on average a one point higher life satisfaction than comparable non-mothers.[11]

The gradual increase in mothers' satisfaction could be the result of positive life events which are conducive to the decision to start a family (marriage, increased household income, etc.). However, the socioeconomic variables in $\mathbf{x}_{it}$ explain surprisingly little of the gap before first birth, as the dotted line shows. This indicates the presence of substantial positive selection on unobservables. If mothers' life satisfaction decreased after transition, this self-selection would lead standard regression approaches to underestimate the effect of motherhood.

Table 4.1 contains regression results which confirm the stylized facts visible from Figure 4.2. The estimates correspond again to model (4.1), but the large number of monthly indicators has been collapsed into three periods: pregnancy, from pregnancy to five years before first birth, and before five years.[12] Mothers and non-mothers start out having virtually the same expected happiness. Some difference is visible in the years before birth. Pregnancy is characterized by large satisfaction gains.[13]

To investigate selection further, we use information on planned and unplanned pregnancies which is available for a subsample of the GSOEP, and replicate Figure 4.2.[14] The vector containing months to first birth is interacted with an indicator whether the pregnancy was planned or not. Figure 4.3 plots the results. Mothers with planned pregnancies

[11] The lines plotted in Figure 4.2 have been smoothed, which makes the effect appear smaller.

[12] The last period goes beyond the limit of seven years shown in Figure 4.2. The earliest observations are up to 20 years before first birth. However, the number of observations diminishes very fast with increasing time to first birth.

[13] We also replicated these estimations using yearly birth data and obtained very similar results.

[14] The women in this subsample are from younger cohorts. For further details refer to Appendix A.4.

– the large majority – exhibit the same increasing trend as before. Mothers with unplanned pregnancies have lower average satisfaction. The path is also more volatile, but this might be a consequence of the small sample size. Up to the pregnancy period, there is little evidence for a trend in their satisfaction. However, the evolution during pregnancy mirrors that of planned motherhoods.

Since the pregnancy effect is present in unplanned motherhoods and similar to that of planned motherhoods, we will treat this "anticipation" as part of the satisfaction gains due to motherhood. In contrast, we view the satisfaction differences in the period five years before first birth up to pregnancy as the result of positive selection on unobservables which we seek to account for directly in our estimations.

## 4.3 Empirical strategies

We propose three different empirical approaches that embed the increase in life satisfaction during the five years prior to first birth. The first two approaches contrast the life satisfaction trajectory of prospective mothers from pregnancy on with the trajectory of a comparable non-mother. These empirical strategies are (i) a nearest neighbor matching estimator that pairs mothers to the most similar non-mothers in terms of pre-birth covariates and pre-birth life satisfaction, and (ii) a regression which controls for pre-birth covariates and the average pre-birth life satisfaction trend and level. Intuitively, both approaches identify the effect of motherhood by comparing future life satisfaction of similar women who experience the same evolution of happiness, but only some of these women become mothers. The third approach does not rely on a comparison between mothers and non-mothers, but exploits intrapersonal variation. A fixed effect regression with dummy variables for the last five pre-birth years is proposed. This strategy estimates the effect of motherhood on life satisfaction by contrasting mothers' life satisfaction after birth to levels reported prior to the five year long satisfaction increase. Whereas all three regression

models differ, all of them preclude self-selection of mothers to affect the estimation of the motherhood effect. The yearly effects can be estimated for the pregnancy period and the first twenty years following birth. While the analysis is restricted to this window owing to the requirement to observe mothers five years before first birth, Figure 4.1 suggested that satisfaction paths of mothers and non-mothers converge in later years anyhow.

### 4.3.1 Nearest neighbor matching

We employ the nearest-neighbor matching estimator with bias correction proposed by Abadie and Imbens (2002; see also Abadie et al., 2004). We match mothers and non-mothers based on age at first birth, values of socioeconomic covariates in the year before birth, and life satisfaction during five, four, three and two years before birth.[15] For instance, consider a hypothetical exact match: A mother with age at first birth 25 is matched to a 25 year old non-mother; both had the same socioeconomic variables at age 24, and both have had the same life satisfaction trajectory from age 20 to 23. Non-mothers can be used to match various ages of first birth. In the previous example, the same non-mother at age 26 can serve as a match to a mother with age at first birth 26. In that case, non-mother's covariates are measured at age 25 and past life satisfaction is measured from age 21 to 24. In practice, there are no exact matches over the whole set of conditioning variables, and we match exactly on past life satisfaction paths while using the four nearest matches in terms of Mahalanobi distance for the remaining variables.[16]

For every age of the first born child $p = -1, 0, 1, 2, \ldots 20$, the matching estimator of the motherhood satisfaction effect reads:

$$\beta_p = \frac{1}{N_p} \sum_{i=1}^{N_p} ls_{ip} - \widehat{ls}_{ip}. \tag{4.2}$$

[15] We use the same socioeconomic variables as before: relationship status, working hours, education, household members, household income. In addition we match on survey wave and years in panel.

[16] Details on the dataset are discussed in Appendix A.5. Mahalanobi distance is the Euclidean distance between all matching variables weighted by their inverse covariance matrix (cf. Abadie and Imbens, 2002). Our results are robust to the use of other number of nearest neighbors, such as the single nearest, two and six nearest neighbors.

The variable $\widehat{ls}_{ip}$ denotes mother $ip$'s predicted life satisfaction if she would not have a child. It equals $\frac{1}{4}\sum_{j\in J_i}^{4} ls_{jp}$, where $J_i$ is the set of the four most similar individuals to mother $i$ from the group of non-mothers. $N_p$ is the number of mothers observed $p$ years after first delivery. Thus, the effect (4.2) can be interpreted as the average treatment effect on the treated for the "treatment" motherhood.

### 4.3.2 Regression using past satisfaction levels and trends

Similar in spirit to the matching estimator, this regression contrasts mothers and non-mothers conditioning on pre-birth satisfaction levels and trends. As before, non-mothers were assigned to all possible ages of first birth in order to determine "pre-birth" realizations of their covariates and "post-birth" satisfaction. The regression equation is

$$ls_{it} = \alpha + m_i \cdot \mathbf{yab}'_{it}\beta + \mathbf{yab}'_{it}\gamma + \theta_1 avg(pls)_i + \theta_2 tr(pls)_i + \mathbf{x}'_{it}\delta + \varepsilon_{it}. \tag{4.3}$$

The variable $m_i$ is an indicator that equals one for mothers and zero for non-mothers. The vector $\mathbf{yab}_{it}$ contains a set of dummy variables for "years after first birth" ranging from -1 to 20. The motherhood variable $m_i$ is interacted with $\mathbf{yab}_{it}$. Thus, mothers' satisfaction path relative to non-mothers during pregnancy and the next twenty years is captured by $\beta$. The variables $avg(pls)_i$ and $tr(pls)_i$ control for pre-birth differences in satisfaction two to five years before birth; $avg(pls)_i$ is the average past life satisfaction level and $tr(pls)_i$ –$tr$ stands for trend– is the average yearly change in satisfaction. The vector $\mathbf{x}_{it}$ contains all socioeconomic covariates one year before birth as well as survey year and number of interviews.[17]

Such an analysis places heavy demands on the data. At least four observations per woman need to be available to be included in the estimation sample; mothers must be surveyed before and after giving birth to their first child.[18]

[17] Robustness checks were performed lagging covariates three and five years, producing virtually no changes in the results.

[18] The resulting dataset is described in Appendix A.5. Replacing average level and average trend with satisfaction lags as in the matching approach reduces the estimation sample further. Our results are robust to such a specification, too.

### 4.3.3 Fixed effect regression accounting for the anticipation effect

In contrast to the first two estimation strategies the fixed effects regression exploits intrapersonal variation only to identify the effect of motherhood. Hence, this approach does not rely on a contrast between two non-randomly selected groups from the population and controls for time-invariant individual-specific unobserved heterogeneity, such as personality traits. We implement the following specification:

$$ls_{it} = \alpha_i + \mathbf{afc}'_{it}\beta + \mathbf{age}'_{it}\gamma + \mathbf{pre}'_{it}\theta + \mathbf{x}'_{it}\delta + \varepsilon_{it} \tag{4.4}$$

The vector $\mathbf{afc}_{it}$ contains a set of dummy variables for "age of first child" ranging from -1 to 20. All elements of $\mathbf{afc}_{it}$ are zero for non-mothers; i.e. non-mothers contribute to the identification of the parameters of other covariates only. The model is similar to the regression with past satisfaction level and trend. However, pre-birth covariates and controls for pre-birth satisfaction paths are missing because parameters of time invariant variables are not identified anymore (reducing $\mathbf{x}_{it}$ to controls for survey year and years in panel). They are absorbed into the fixed effects $\alpha_i$. In order to account for the heightened levels of satisfaction during the five years preceeding birth, i.e. to avoid overestimation of individual fixed effects, a set of four dummy variables is included in the regression ($\mathbf{pre}_{it}$), indicating each of mothers' four years of the anticipation period before pregnancy.[19]

Out of the three regression models, the fixed effect regression is the least demanding on data. All observations, no matter how long in the sample and whether observed before or after birth can be used to identify at least part of the motherhood effect's dynamics, resulting in a visibly increased sample size.[20]

[19] For non-mothers, all elements of $\mathbf{pre}_{it}$ are equal to zero.
[20] The data is detailed in Appendix A.6.

## 4.4 Results

### 4.4.1 Main results

Figure 4.4 shows the estimated effects of motherhood for the year before birth of the first child and for the following twenty years. The figure presents results for the three approaches discussed in Section 4.3. The solid line depicts the results of the fixed effects estimation. The dashed and the dotted line, show the results of the regression with past satisfaction level and trend, and the results of the matching approach. An effect in the order of one third point, for example, five years after first child's birth, describes an average life satisfaction difference between mothers and non-mothers of 0.3 points on the 11 points scale. The point estimates used to produce the graph, the corresponding standard errors, and more details on the regressions can be found in Table B.1 (Appendix B).

All three strategies lead to strikingly similar results, especially in the first years after delivery. The figure shows that prospective mothers are happier compared to non-mothers one year before childbirth. The maximum life satisfaction difference between mothers and non-mothers is reached in the year of delivery. The effect is then over half a satisfaction point. The point estimates lie between 0.52 and 0.56 (see Table B.1). This is a substantial effect compared to the influence of other standard variables in happiness regressions like income or age. The difference in life satisfaction between mothers and non-mothers diminishes with age of the first born child, a sign of adaptation. However, the effect remains positive over the first twenty years of motherhood. The hypothesis that motherhood has no effect on life satisfaction, thus that all shown coefficients are equal to zero, is clearly rejected by an F-test (see Table B.1). However, even in the fixed effects regression, which gives the most precise estimates, only the coefficients capturing the effects during the year of birth and one year before and after birth are individually significant at the 5% level. The imprecise estimates, evoked by the small number of women who are observed before and some time after childbirth, are also the most likely explanation why the point estimates

of the different approaches slightly diverge in late years. Against the picture drawn in previous studies, these results suggest that once mothers are compared to ex-ante similar non-mothers, motherhood affects life satisfaction positively.

### 4.4.2 Comparison to previous approaches

Previous studies which looked at the association between children and life satisfaction have found mostly a negligible or negative motherhood effect. To see whether our results are driven by our special sample restrictions or by the different identification strategy, we replicate regressions as they are typically found in the literature with the samples used in this study. Thus, motherhood is identified through a dummy variable indicating the presence of at least one child in the household; and contemporaneous realizations for all control variables are employed. For all samples a regression with and without fixed effects is estimated. Table 4.2 reports the results from estimating such a life satisfaction model. The first two columns with heading "Transition sample" contain the estimates for the sample which was used for the matching approach and the regression with controls for past satisfaction. Column three and four ("FE sample") present the results with observations used in the fixed effects regression. The last two columns ("GSOEP") present results using all women that have participated at least once in the GSOEP.

Five out of six estimates are negative and all of them are insignificant, regardless whether fixed effects are included or not. Thus, the standard approach is unable to detect the positive effects of motherhood clearly present when comparing life satisfaction paths of mothers to that of ex-ante similar non-mothers.

### 4.4.3 Extensions

We extend our analysis in different directions. First, we examine whether mother's age of first birth affects satisfaction gains obtained from motherhood. Then, we study if motherhood status captures the main effect of the fertility decision on life satisfaction or if one

should focus on the number of children. Finally, we explore the effect of fatherhood on life satisfaction. Except where noted otherwise, we use the fixed effect specification in this section.

### Age at first birth

Figure 4.5 shows the effect of motherhood on life satisfaction depending on mother's age at first birth (AFB in the figure). For comparison, the thick line depicts again the average effect for all mothers presented earlier in Figure 4.5. The effects for different groups of age at first birth are shown by the thin lines. The youngest group, for example, consists of mothers giving birth to their first child between the age 26 and 29. Looking at younger mothers is difficult, because six pre-birth observations are needed to allow for individual-specific fixed effects and an anticipation period of five years. The oldest group consists of women with first delivery between 35 and 37. The different group lines are smoothed to present a visually clearer picture.

The horizontal order of the four lines suggests that the motherhood effect is larger for women having a child later in life.[21] The lines of the two younger groups are below the average line and the curves for the two older groups above. The oldest category have clearly the largest happiness gains. The youngest mothers, on the other hand, seem to be the only group of mothers that suffer from the motherhood status, at least in later years. Since the pregnancy effect seems higher for older groups than for younger groups, one has to be cautious with interpreting the results. If only the difference in the happiness levels directly before and after delivery is considered, the women in the oldest category still profit most and the youngest mothers fewest, but the ranking of the middle groups is less clear.

[21] There are several possible channels which might explain such a pattern. For instance, later timing of first birth is associated with higher wage growth (Herr, 2007).

### Single-child and multiple-parity mothers

Figure 4.6 shows the effect of motherhood on life satisfaction for single child mothers and mothers giving birth to several children in the observation period. Effects for both groups of mothers are strikingly similar a year around childbirth. The differences in life satisfaction levels between the two categories of mothers and non-mothers are small from five years after delivery on. In between, however, multiple-party mothers report higher happiness levels on average. The reason is probably the additional birth taking place during this period. We looked also at the effect of the second child, and the results (not shown) support this interpretation. In about seventy percent of all cases, the time span between birth of the first and second child amounts to four years or less, and the effect of the second child is also positive with a peak at childbirth, albeit the effect is only about half as large as the effect caused by the first child's birth. All in all, these results suggest that the main event or decision in a life of a mother is birth of the first child and the related issue of starting a family. The intensive margin of fertility, number of children, seems less important for the overall evolution of mothers' life satisfaction paths.

### Fatherhood

Fatherhood has been left out so far for two reasons. First, identification of fathers identity in the data is far less reliable than mothers. The GSOEP is a household survey and fathers may often not share the same household. Thus, direct pointers are often missing. Second, it is more difficult to define an appropriate age threshold for defining men's completed fertility as their distribution of age at birth exhibits a noticeably longer tail than women's. With these shortcomings in mind, we replicated the estimations for fathers. Again the empirical distribution of age at first birth was used to determine the maximum age at first birth (47 years).[22]

[22] Until the age of 48, 99.8% of fathers have had their first child. Appendix A.7 depicts the estimation sample in detail.

Figure 4.7 shows the effect of fatherhood. The results are similar to those of motherhood, however the effect before and at birth seem a bit smaller. Whereas the effect of motherhood in the first year after birth was estimated to be about 0.55 points, the effect of fatherhood is about 0.45. The fixed effects estimator shows a clear decline after two years, stabilizing around 0.1 for the next twenty years; while the matching estimator and the regression with past satisfaction level and trend suggest a slower decline. Thus, both men and women seem to benefit from having a child.

## 4.5 Conclusions

This paper has presented evidence of self-selection into motherhood and proposed approaches to estimate satisfaction gains of parenthood which account for the positive selection. This is a sharp contrast to the usual analysis in the literature, which relies on *ex-post* comparisons between parents and non-parents and uses observations of prospective parents as part of the control group. We overcome the censoring of potential mothers by the construction of a completed fertility decision sample. Moreover, we find evidence for self-selection into motherhood and account for it in our analyses by using *ex-ante* information on observables and on previous satisfaction paths. The results are robust to the various specifications and consequently confirm the importance to factor selection issues in. Moreover, our estimates contrast with those of the previous literature in that we uncover a positive effect of motherhood - a finding which is in line with a mainstream view of choice behavior based on utility maximization.

The motherhood effect can be put into pecuniary terms. With knowledge of the discount factor in the intertemporal utility function it is possible, in principle, to compute the equivalent amount of household income which makes women indifferent between motherhood and childlessness. We use discount factors of 0.9 and 0.8 to calculate the net present value of motherhood. Estimates of discount factors found in the literature vary

considerably (Frederick, Loewenstein and O'Donoghue, 2002). Our first discount factor lies approximately in the middle of the range reported in recent field studies. Discount factors obtained experimentally are typically higher, which is reflected in the second choice. We monetize the yearly satisfaction differentials for mothers (by comparing the respective motherhood coefficient to the coefficients on income) and then discount them to the year before pregnancy using estimates of our specifications with FE and with lags. Based on the FE results, for the median woman motherhood is worth about 1.2 net yearly household incomes using the stronger discount rate, and about 1.7 using the weaker one. Using the results of the regressions with lags, the compensating variation is about 1.1 or 1.9 yearly incomes (based on discount factors 0.8 and 0.9, respectively). These estimates seem reasonable. For instance, couples' willingness to pay for expensive assisted fertility treatments suggest that expected utility gains from motherhood need to be substantial.[23] Another indication of children's high value to parents, happiness losses caused by the death of a child have been valued at similarly high magnitudes (Oswald and Powdthavee, 2008).

Obviously, the utility gains from motherhood are specific to social, technological and other factors. The women surveyed in the German Socio-Economic Panel live in a modern society and a historical moment where birth control is effective, widely available and its use socially accepted; there is universal health care access and the law stipulates extended maternity leaves. Thus, such an environment is probably particularly conducive to large satisfaction gains from motherhood.

[23] Cost-effectiveness studies estimate the cost of live birth at about USD 50,000 (in year 2002 prices; cf. Collins, 2002). In Germany, a part of assisted fertility treatment costs are covered by health insurance. However, there are substantial further non-pecuniary costs such as emotional stress and health risks associated with assisted fertility treatments (Gumus and Lee, 2012).

# References


Abadie, A. and G. W. Imbens, 2002, "Simple and bias-corrected matching estimators for average treatment effects", *NBER technical working paper*, 283

Abadie, A., D. Drukker, J. L. Herr and G. W. Imbens, 2004, "Implementing matching estimators for average treatment effects in Stata", *Stata Journal*, 4, 290–311

Alesina, A., R. Di Tella and R. MacCulloch, 2004, "Inequality and happiness: are Europeans and Americans different?", *Journal of Public Economics*, 88, 2009–2042

Angrist, J. D. and W. N. Evans, 1998, "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size", *The American Economic Review*, 88, 450–477

Arroyo, C. R. and J. Zhang, 1997, "Dynamic microeconomic models of fertility choice: A survey", *Journal of Population Economics*, 10, 23–65

Becker, G., 1960, "An economic analysis of fertility", In Becker, G. (Ed.), *Demographic and Economic Change in Developed Countries*, Princeton, NJ: Princeton University Press

Benjamin, D. J., O. Heffetz, M. S. Kimball and A. Rees-Jones, 2012, "What Do You Think Would Make You Happier? What Do You Think You Would Choose?", *American Economic Review*, 102, 2083–2110

Blanchflower, D. G., 2009, "International Evidence on Well-Being", In A. B. Krueger (Ed.), *Measuring the Subjective Well-Being of Nations: National Accounts of Time Use and Well-Being*, 155–226, Chicago, IL: University of Chicago Press

Clark, A. E., 2007, "Born To Be Mild? Cohort Effects Don't (Fully) Explain Why Well-Being Is U-Shaped in Age", *IZA Discussion Papers*, 3170

Clark, A. E, E. Diener, Y. Georgellis and R. E. Lucas, 2008, "Lags and leads in life satisfaction: a test of the baseline hypothesis", *Economic Journal*, 118, 529, F222–F243

Clark, A. E, P. Frijters and M. A. Shields, 2008, "Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles", *Journal of Economic Literature*, 46, 95–144

Clark, A. E. and A. J. Oswald, 1994, "Unhappiness and unemployment", *Economic Journal*, 104, 648–59

Clark, A. E. and A. J. Oswald, 2002, "Well-being in panels", mimeo, University of Warwick, UK

Collins, J. A., 2002, "An international survey of the health economics of IVF and ICSI", *Human Reproduction Update*, 8, 265–277

Del Boca, D. and M. Locatelli, 2006, "The Determinants of Motherhood and Work Status: A Survey", *IZA Discussion Papers*, 2414

Del Bono, E., A. Weber and R. Winter-Ebmer, 2012, "Clash of career and family: Fertility decisions after job displacement", *Journal of the European Economic Association*, 10, 659–683

Di Tella, R., R. J. MacCulloch and A. J. Oswald, 2001, "Preferences over inflation and unemployment: Evidence from surveys of happiness", *American Economic Review*, 91, 335–341

Di Tella, R., R. J. MacCulloch and A. J. Oswald, 2003, "The macroeconomics of happiness", *Review of Economics and Statistics*, 85, 809–827

Dolan P., T. Peasgood and M. White, 2008, "Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being", *Journal of Economic Psychology*, 29, 94–122

Easterlin, R. A., 1973, "Does Money Buy Happiness?", *The Public Interest*, 30, 3–10

Easterlin, R. A., 2001, "Income and Happiness: Towards an Unified Theory", *Economic Journal*, 111, 465–84

Ferrer-i-Carbonell, A., 2012, "Happiness Economics", forthcoming in: *SERIEs: Journal of the Spanish Economic Association*, published online February 24 2012, DOI 10.1007/s13209-012-0086-7

Frederick, S., G. Loewenstein and T. O'Donoghue, 2002, "Time Discounting and Time Preference: A Critical Review", *Journal of Economic Literature*, 40, 351–401

Frey, B. S. and A. Stutzer, 2002, "What can economists learn from happiness research?", *Journal of Economic Literature*, 40, 402–435

Frijters, P., D. W. Johnston and M. Shields, 2011, "Happiness dynamics with quarterly life event data", *Scandinavian Journal of Economics*, 113, 190–211

Gilbert, D. P., 2006, "Stumbling on Happiness", London: Harper Perennial

Gumus, G. and J. Lee, 2012, "Alternative paths to parenthood: IVF or child adoption?", *Economic Inquiry*, 50, 802–820

Hansen, T., 2011, "Parenthood and happiness: A review of folk theories versus empirical evidence", *Social Indicators Research*, 123, 1–36

Herbst, C. M. and J. Ifcher, 2012, "A bundle of joy: does parenting really make us miserable?", *SSRN Working Paper*, No. 1883839 (Version: May 16, 2012)

Herr, J. L., 2007, "Does it Pay to Delay? Understanding the Effect of First Birth Timing on Women's Wage Growth", unpublished manuscript, University of California, Berkeley

Kahneman, D., E. Diener and N. Schwarz (Eds.), 1999, *Well-Being: The Foundations of Hedonic Psychology*, New York: Russell Sage Foundation

Kohler, H.-P., J. R. Behrman and A. Skytthe, 2005, "Partner + Children = Happiness? The Effects of Partnerships and Fertility on Well-Being", *Population and Development Review*, 31, 407–445

Layard, R., 2005, "Happiness: Lessons from a New Science", London: Allen Lane

Ma, B., 2010, "The Occupation, Marriage, and Fertility Choices of Women: A Life-Cycle Model", *UMBC Economics Department Working Papers*, 10-123

Michaud, P.-C. and K. Tatsiramos, 2011, "Fertility and female employment dynamics in Europe: the effect of using alternative econometric modeling assumptions", *Journal of Applied Econometrics*, 26, 641–668

Morgan, P. and R. Berkowitz King, 2001, "Why have children in the 21st century? Biological predisposition, social coercion, rational choice", *European Journal of Population*, 17, 3–20

Oswald, A. J. and N. Powdthavee, 2008, "Death, Happiness, and the Calculation of Compensatory Damages", *Journal of Legal Studies*, 37, S217–S251

Stanca, L., 2012, "Suffer the little children: Measuring the effects of parenthood on well-being worldwide", *Journal of Economic Behavior and Organization*, 81, 742–750

Stevenson, B. and J. Wolfers, 2008, "Economic growth and subjective well-being: Reassessing the Easterlin paradox", *NBER Working Paper Series*, No. 14282

Stutzer, A. and B. S. Frey, 2006, "Does marriage make people happy, or do happy people get married?", *Journal of Socio-Economics*, 35, 326–347

Van Landeghem, B., 2012, "A test for the convexity of human well-being over the life cycle: Longitudinal evidence from a 20-year panel", *Journal of Economic Behavior and Organization*, 81, 571–582

Vanassche, S., G. Swicegood and K. Matthijs, 2012, "Marriage and Children as a Key to Happiness? Cross-National Differences in the Effects of Marital Status and Children on Well-Being", forthcoming in *Journal of Happiness Studies*, published online May 1 2012, DOI 10.1007/s10902-012-9340-8

Veenhoven, R., 2008, "Healthy happiness: effects of happiness on physical health and the consequences for preventive health care", *Journal of Happiness Studies*, 9, 449–469

Wilde, E. T., L. Batchelder and D. T. Ellwood, 2010, "The Mommy Track Divides: The Impact of Childbearing on Wages of Women of Differing Skill Levels", *NBER Working Paper Series*, No. 16582

Willis, R. J., 1973, "A New Approach to the Economic Theory of Fertility Behavior", *Journal of Political Economy*, 81, S14–S64

Winkelmann, L. and R. Winkelmann, 1997, "Why are the unemployed so unhappy? Evidence from panel data", *Economica*, 65, 1–15

Wunder, C., A. Wiencierz, J. Schwarze and H. Kuechenhoff, 2011, "Well-being over the life span: semiparametric evidence from British and German longitudinal data", forthcoming in *Review of Economics and Statistics*, published online July 19 2011, DOI 10.1162/REST_a_00222

Table 4.1: OLS estimates of satisfaction differences between prospective mothers and non-mothers

| | (1) | (2) |
|---|---|---|
| Pregnancy (9 months to 1 month before birth) | 0.71 | 0.65 |
| | (0.13) | (0.13) |
| 5 years to 10 months before birth | 0.23 | 0.16 |
| | (0.13) | (0.12) |
| More than 5 years before birth | 0.01 | 0.03 |
| | (0.17) | (0.16) |
| Socioeconomic control variables | No | Yes |
| Number of observations | 5,756 | |
| Number of individuals | 947 | |

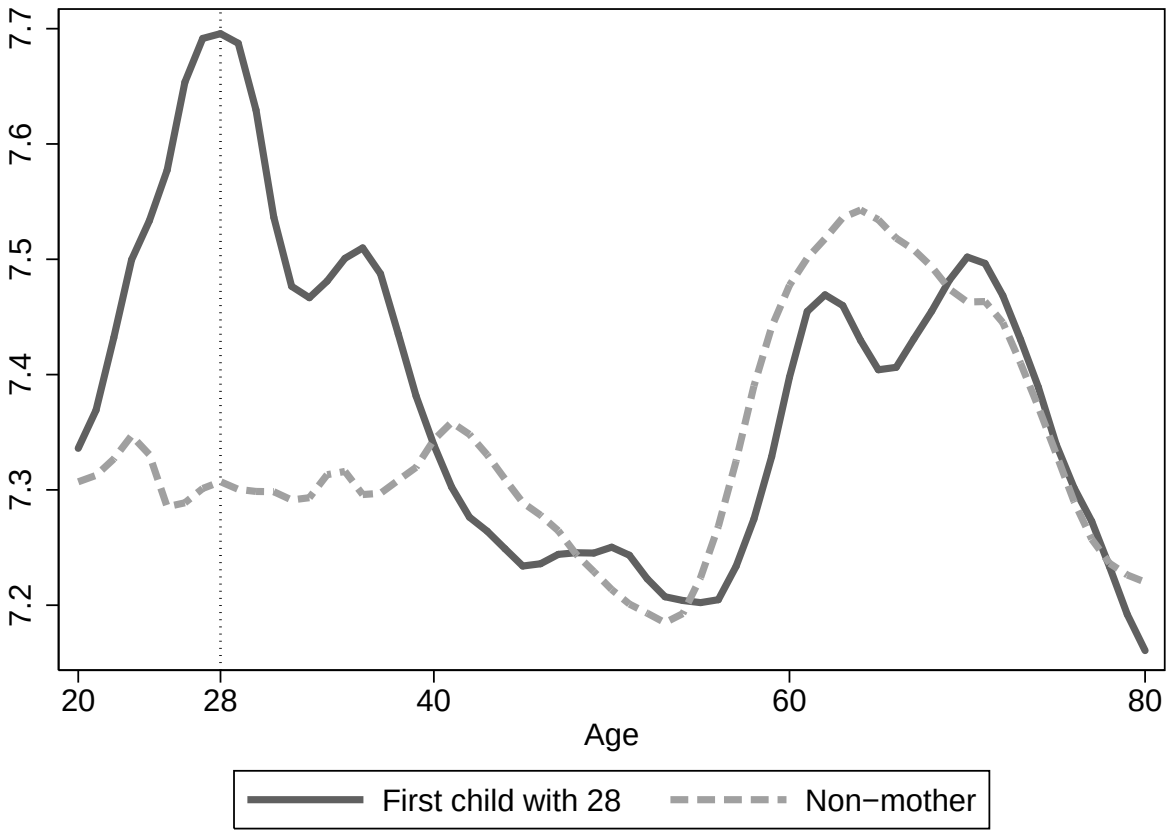
*Notes:* Cluster robust standard errors in parentheses. Both regressions include full sets of age dummies and of number of years in panel. The regression in column (2) additionally includes the following control variables: married, boyfriend, single, second order polynomials of weekly working hours and household income and full sets of dummies for education and number of household members.

Table 4.2: Estimates of satisfaction gains of motherhood using standard approaches from the literature

| | Transition sample | | FE sample | | GSOEP | |
|---|---|---|---|---|---|---|
| Child dummy | -0.036 | -0.104 | -0.004 | -0.015 | 0.028 | -0.044 |
| | (0.091) | (0.069) | (0.052) | (0.043) | (0.035) | (0.030) |
| Individual FE | No | Yes | No | Yes | No | Yes |
| Number of obs. | 25,910 | | 78,470 | | 198,016 | |
| Number of individuals | 1,590 | | 9,791 | | 22,510 | |

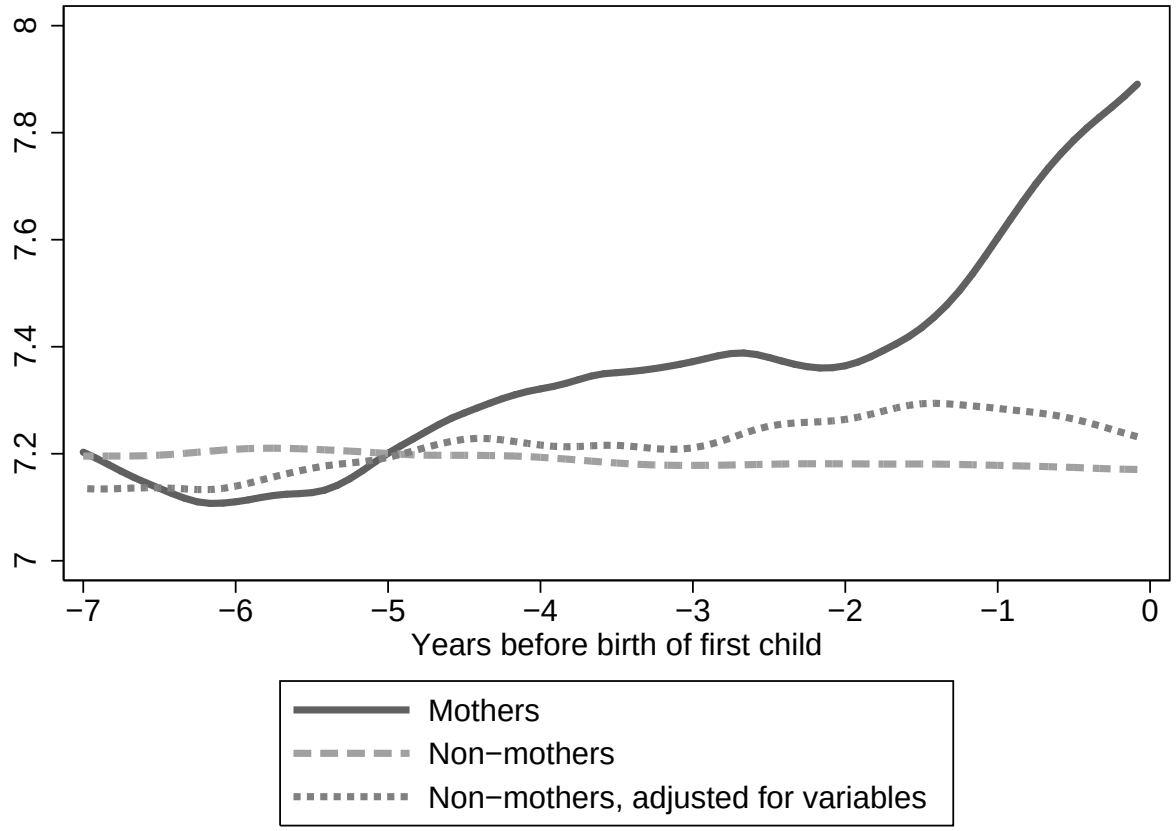
*Notes:* Cluster robust standard errors in parentheses. The regressions additionally include the following control variables: married, boyfriend, single, second order polynomials of weekly working hours and household income and full sets of dummies for age, education, number of household members and years in panel.

Figure 4.1: Life satisfaction of women over the life cycle



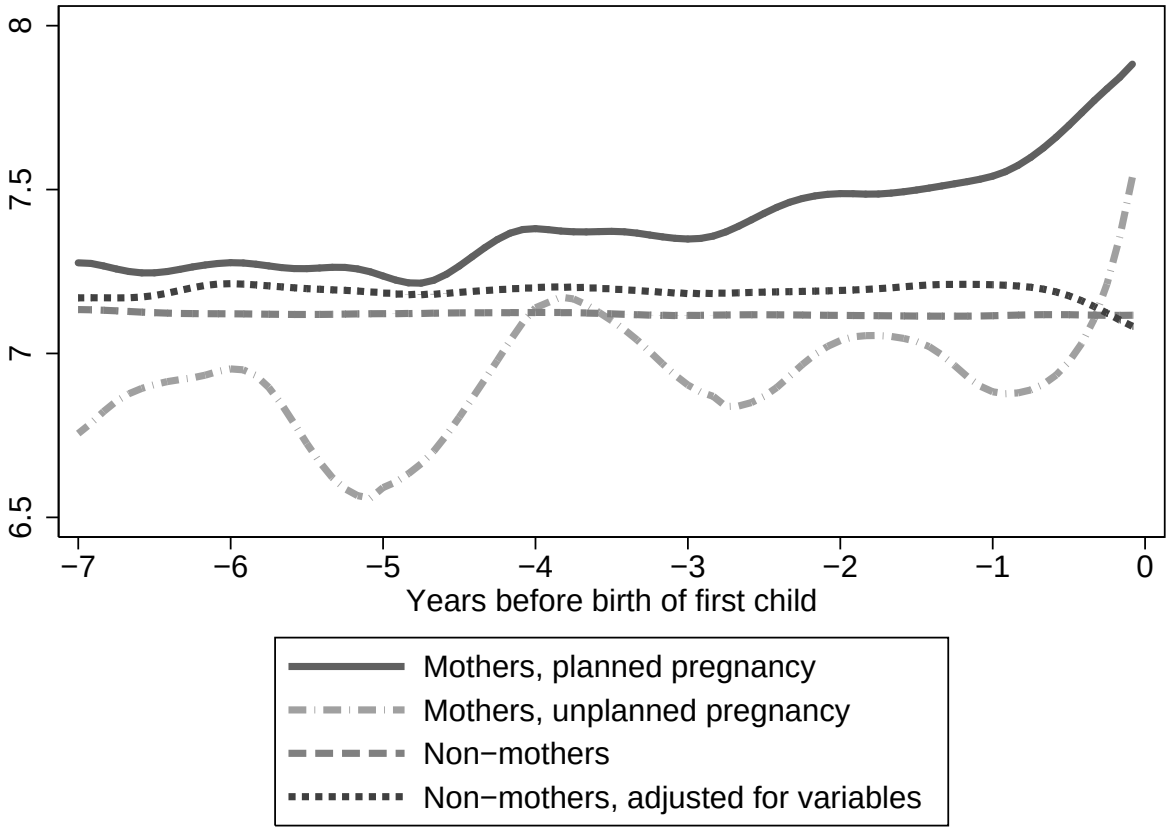

*Notes:* Data from the GSOEP waves 1984-2009 is detailed in Appendix A.2. Displayed average life satisfaction paths are conditional on sets of dummies for survey years and years in panel, smoothed (Lowess) with bandwidth 0.12.

Figure 4.2: Life satisfaction before birth



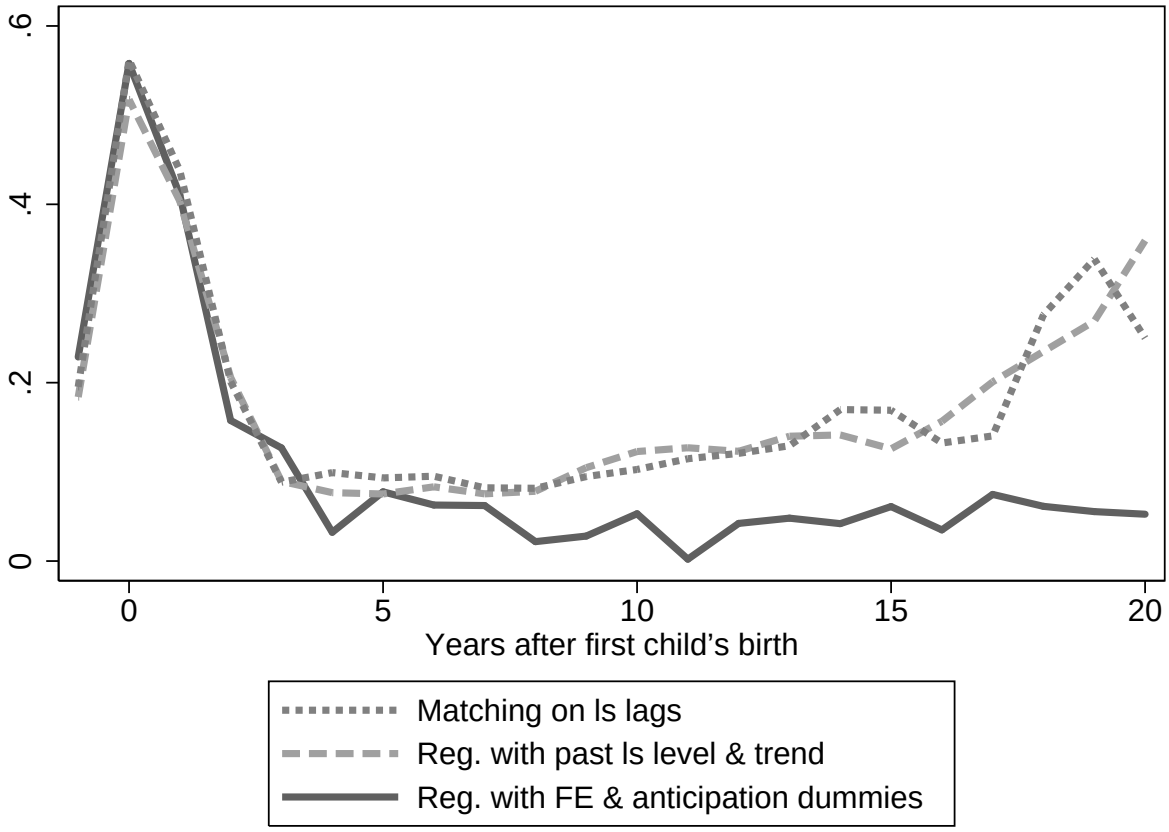

*Notes:* The graph depicts parameter estimates for the variable **months to birth** in model (4.1) for a subset of 7 years. The data is detailed in Appendix A.3. Displayed average life satisfaction paths are conditional on sets of dummies for survey years and years in panel. Predicted life satisfaction adjusted for variables further includes controls for education, relationship status, household members, working hours and household income. All lines smoothed (Lowess) with bandwidth 0.3 Appendix A.1 contains a detailed description of the included terms.

Figure 4.3: Life satisfaction before birth - Planned vs. unplanned pregnancies



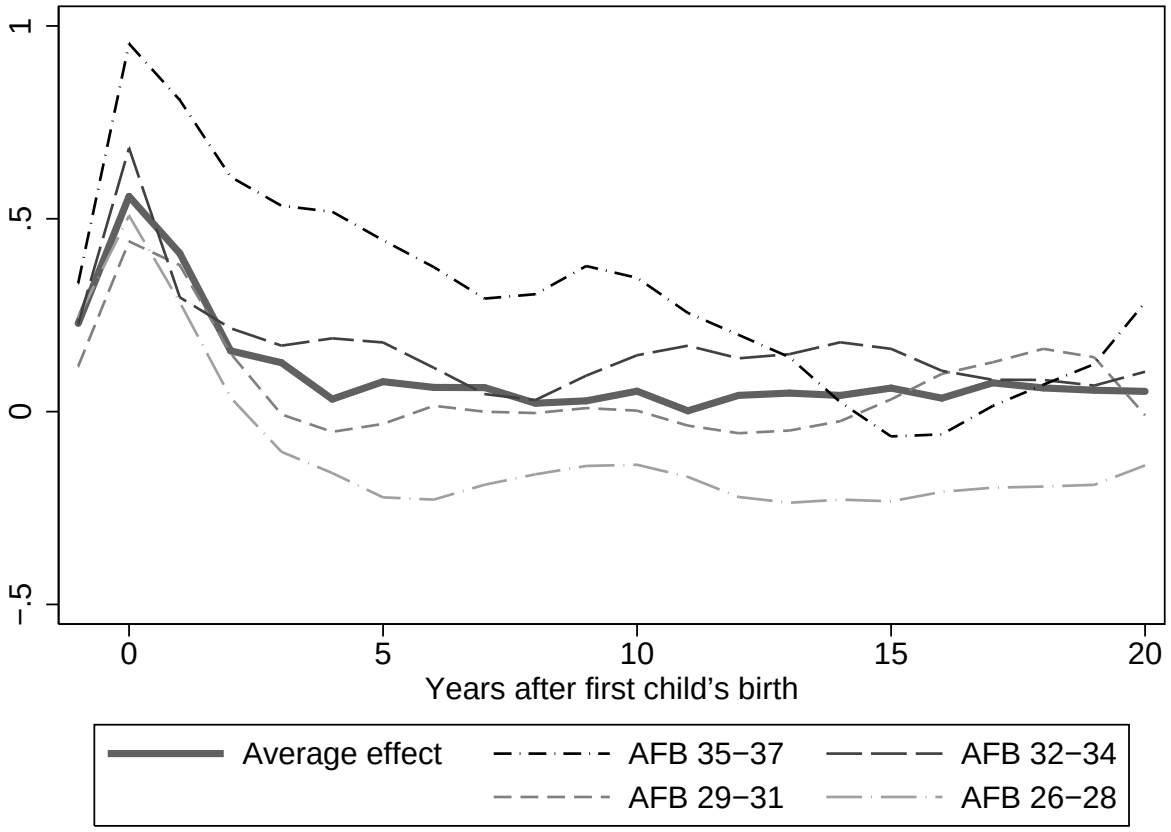

*Notes:* The graph depicts parameter estimates for the variable **months to birth** in model (4.1) interacted with a dummy indicating whether motherhood was planned or not, for a subset of 7 years. The data is detailed in Appendix A.4. Displayed average life satisfaction paths are conditional on sets of dummies for survey years and years in panel. Predicted life satisfaction adjusted for variables further includes controls for education, relationship status, household members, working hours and household income. All lines smoothed (Lowess) with bandwidth 0.3. Appendix A.1 contains a detailed description of the included terms.

Figure 4.4: Estimated life satisfaction (ls) gains of motherhood for different empirical strategies



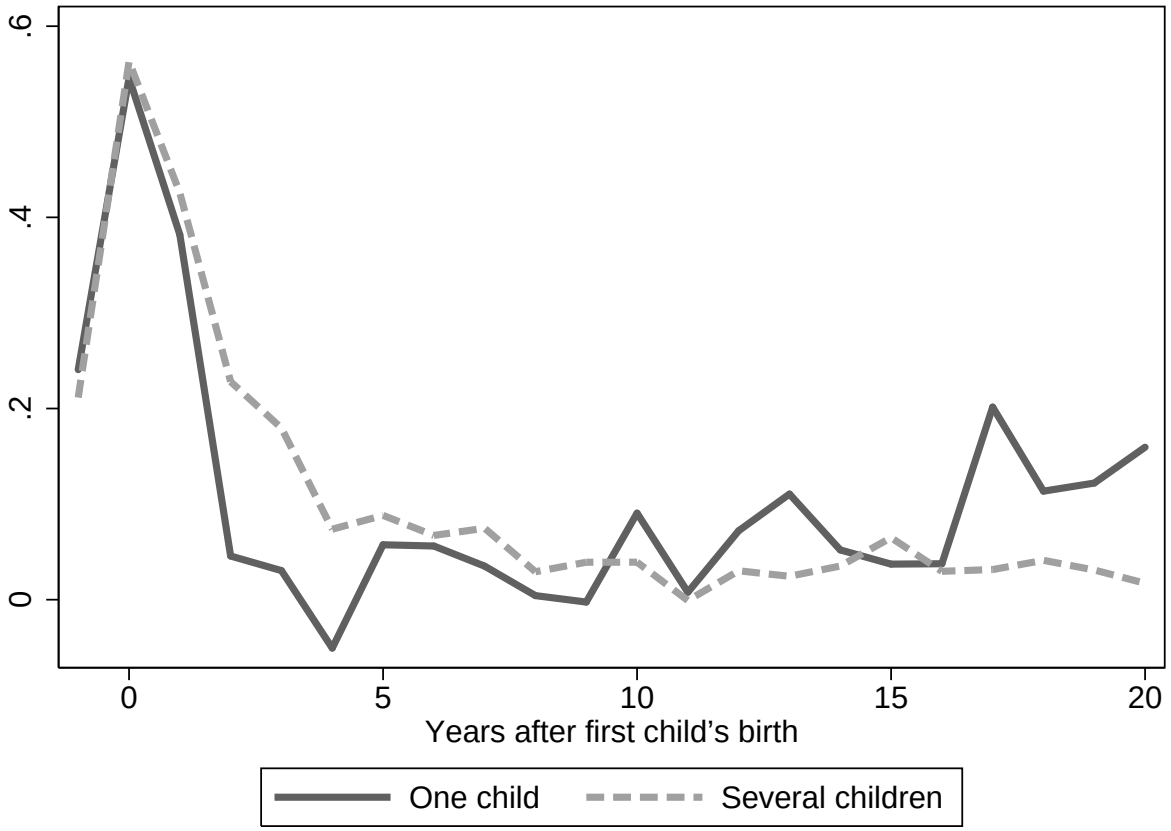

*Notes:* Matching estimates correspond to $\beta_p$ in model (4.2) using the data detailed in Appendix A.5. Matching is achieved on past satisfaction levels from minus two to minus five years and other lagged covariates. Regression with past life satisfaction level and trend correspond to the estimates of $\beta$ in model (4.3). The regression uses the same data as the matching approach. It controls, beside other covariates, for average happiness level two to four years before delivery and the change in the happiness level in the same period. Fixed effect estimation correspond to $\beta$ in model (4.4). It includes four extra dummies for minus two to minus five years before first birth and employs the data introduced in Appendix A.6. All lines smoothed (Lowess) with bandwidth 0.15.

Figure 4.5: Estimated life satisfaction gains of motherhood for different age-at-first-birth (AFB) groups – Fixed effect regression



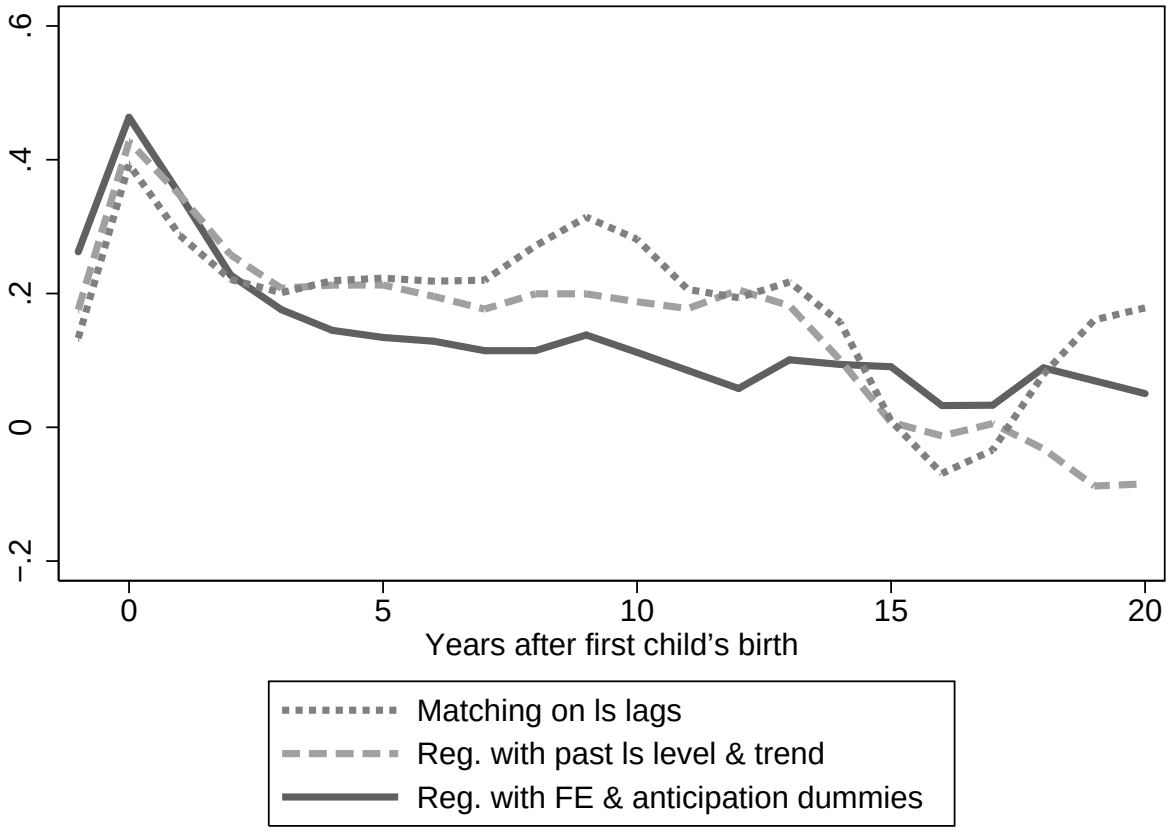

*Notes:* The thick line shows again the average motherhood effect ($\beta$ in model 4.4). The thin lines show the estimated motherhood effect of model (4.4) interacted with age of first birth. All regressions include four extra dummies for minus two to minus five years before first birth. The data is introduced in Appendix A.6. All lines smoothed (Lowess) with bandwidth 0.15.

Figure 4.6: Estimated life satisfaction gains of motherhood for single-child and multiple-parity mothers – Fixed effect regression




*Notes:* The lines show the estimated motherhood effect of model (4.4) interacted with a variable indicating if the mother has one child, or more than one child over her life span. All regressions includes four extra dummies for minus two to minus five years before first birth. The data is introduced in Appendix A.6.

Figure 4.7: Estimated life satisfaction (ls) gains of fatherhood for different empirical strategies




*Notes:* The lines show the fatherhood effect estimated with different approaches. Notes to estimation approaches can be found in Figure 4.4. The data is introduced in Appendix A.7. All lines smoothed (Lowess) with bandwidth 0.15.

# A Data

We use data from German Socio Economic Panel (GSOEP). The GSOEP exhibits at least three features that benefit the analysis of motherhood. First, person pointers identify a respondent's mother and children. Second, we have access to 25 yearly waves, starting in 1984. This permits us to identify women with a fertility rate equal to zero over their entire life, but to observe these non-mothers during possibly fertile years. Third, information on the type of pregnancy (planned or unplanned) is available from a special mother and child questionnaire for the subset of mothers with year of first birth 2002 or later.

Appendix A0 shortly documents how different variables were constructed and how they were integrated as control variables in the regressions. Appendices A1 to A6 describe the subsamples generated from the GSOEP for this study's analyses. Means of selected variables are depicted in Table A.1.

## A.1 Variables used

Original variable names as they appear the first time in the GSOEP are reported in parentheses. Household (ahhnr) and never changing person (persnr) numbers identify households and individuals. Pointers to person numbers define a respondent's mother (mnr, akmutti, bymnr or persnrm), father (byvnr, vnr) and children (kidpnr or idperschild). The dependent variable, life satisfaction, was assessed by asking respondends: "*In conclusion, we would like to ask you about your satisfaction with your life in general. Please answer according to the following scale: 0 means completely dissatisfied, 10 means completely satisfied. How satisfied are you with your life, all things considered?*" (p1110184). Birth year (gebjahr) was used together with survey year to construct age. Exact ages of a mothers' children were computed through birth dates of a child (kidmon, kidgeb) and interview dates of a mother (bpmonin, ahtagin). Years in panel was generated from the number of a respondents' observations in our data.

In all estimations presented in this study, complete sets of indicator variables control for age, survey year and number of years in panel. Estimates controlling for socioeconomic include the following set of variables: seven dummies categories of completed education (apsbil) (secondary school degree, intermediate school degree, technical school degree, upper secondary degree, other degree, dropout, no school degree yet); three dummies for relationship (ap58) married, boyfriend, single; complete set of dummies for numbers of household members (ahhgr); a second order polynomial for weekly hours worked (atatzeit) that range from 0 to 80; a dummy indicating whether hours were reported (58%) or not; household income (hinc84) and household income squared for monthly salaries between 0 and 100,000 Euros and a dummy for reported household income (95%). Moreover, for the pre-birth period analysis the dummy variable planned pregnancy (bcssplan) is used.

## A.2 Life cycle sample

The life cycle analyses include all observations on non-mothers with a fertility rate of zero at age 40 and on mothers with age of first birth equal to 28 years, aged 20 to 80 during waves 1984 to 2009 and reporting valid answers to the questions in this study. This yields 25,773 observations for 3,885 women.

## A.3 Pre-birth completed fertility sample

The pre-birth analysis contrasts pre-birth life satisfaction of mothers-to-be to that of similar non-mothers. Given a threshold of 40 years for a completed fertility decision by the age of 40, prospective mothers are younger than 41 years. This maximum age is imposed on non-mothers' ages, too. This implies that non-mothers are born before 1968. In return, this cohort restriction is applied to mothers' birth cohorts. Moreover, for pre-birth analyses exact ages of respondents' offspring were used. These restrictions leave 5,756 observations for 947 women.

## A.4 Pre-birth birth-type sample

The GSOEP mother and child questionnaire is in field since 2003 and covers new mothers from 2002 on. Out of 1,249 new mothers who answered the question, 70% judged that their pregnancy was more planned than unplanned. Due to the questionnaire's inception date, the information is available for mothers aged maximally 46 years in 2009. To obtain a same-aged control group, the completed fertility decision sample's non-mothers are replaced by potential non-mothers, i.e. contemporaneously childless women. In order to find the same range of age for both mothers and non-mothers, we impose potential non-mothers not to be born before 1959 and not to exceed the age of 40. This leaves us with 14,879 observations for 2,572 individuals. For all of these women first child's exact birth date are available.

## A.5 Transition sample

Implications of matching or controlling on pre-birth life satisfaction are threefold. First, transition into motherhood needs to be observed. This implies that mothers' age cannot exceed 60 years in our sample. We apply this age restriction also to non-mothers. Second, pre-birth observations need to be observed such that controlling or matching on past life satisfaction paths is feasible. For 1,590 women with 25,910 observations past satisfaction levels and trends are identified. Third, our analyses considers mothers one year before first child's birth. To find similar, same-aged non-mothers we use all possible ages of non-mothers. This implies that, if possible, non-mothers are "cloned" and used multiple times with covariates measured at the corresponding age. The total number of observations is then 37,616. Cloning induces an obvious dependence between cloned observations. All reported standard errors and test statistics account for arbitrary clustering and heteroskedasticity of any type at the individual level, and therefore account for the dependence between multiple observations of non-mothers.

## A.6 Fixed effect estimation sample

Fixed effect regressions estimate the effect of motherhood for women aged 20 to 60. The GSOEP provides information about 13,652 women whose ages fall into this interval. Again, only women with a completed fertility decision are retained in the sample. We are left with 78,470 observations for 9,791 individuals.

## A.7 Father sample

For the analysis of fatherhood valid responses of male participants from GSOEP waves 1984 to 2009 are used. As for women, the age by which the fertility decision is completed is defined by means of the data at hand. Mean and median age of first birth for men are equal to 27 and 28 years. 99.6% of all fathers had their first child before the age of 48. We thus define non-fathers as men who have not fathered a child until the age of 48. The sample consists of 82,261 observations for 8,449 men.

Table A.1: Means of selected variables for different samples

| | A1 | A2 | A3 | A4 | A5 | A6 |
|---|---|---|---|---|---|---|
| Proportion parents | 0.35 | 0.52 | 0.35 | 0.81 | 0.90 | 0.93 |
| Age | 51.86 | 30.45 | 27.58 | 31.09 | 34.87 | 39.48 |
| Net-monthly HH-income in Euros | 2137.08 | 2002.23 | 2048.06 | 2200.68 | 2235.64 | 2439.54 |
| Weekly hours worked | 17.64 | 33.69 | 31.10 | 23.38 | 20.15 | 38.06 |
| Proportion high school degrees | 0.19 | 0.25 | 0.33 | 0.23 | 0.18 | 0.19 |
| Proportion school drop outs | 0.03 | 0.02 | 0.02 | 0.03 | 0.04 | 0.03 |
| Proportion married | 0.55 | 0.34 | 0.22 | 0.49 | 0.63 | 0.69 |
| Proportion with partner | 0.16 | 0.36 | 0.54 | 0.29 | 0.18 | 0.15 |
| Proportion single | 0.22 | 0.16 | 0.22 | 0.14 | 0.09 | 0.08 |
| Number of observations | 25,773 | 5,756 | 14,879 | 25,910 | 78,470 | 82,261 |
| Number of individuals | 3,885 | 947 | 2,572 | 1,590 | 9,791 | 8,449 |

# B Regression output

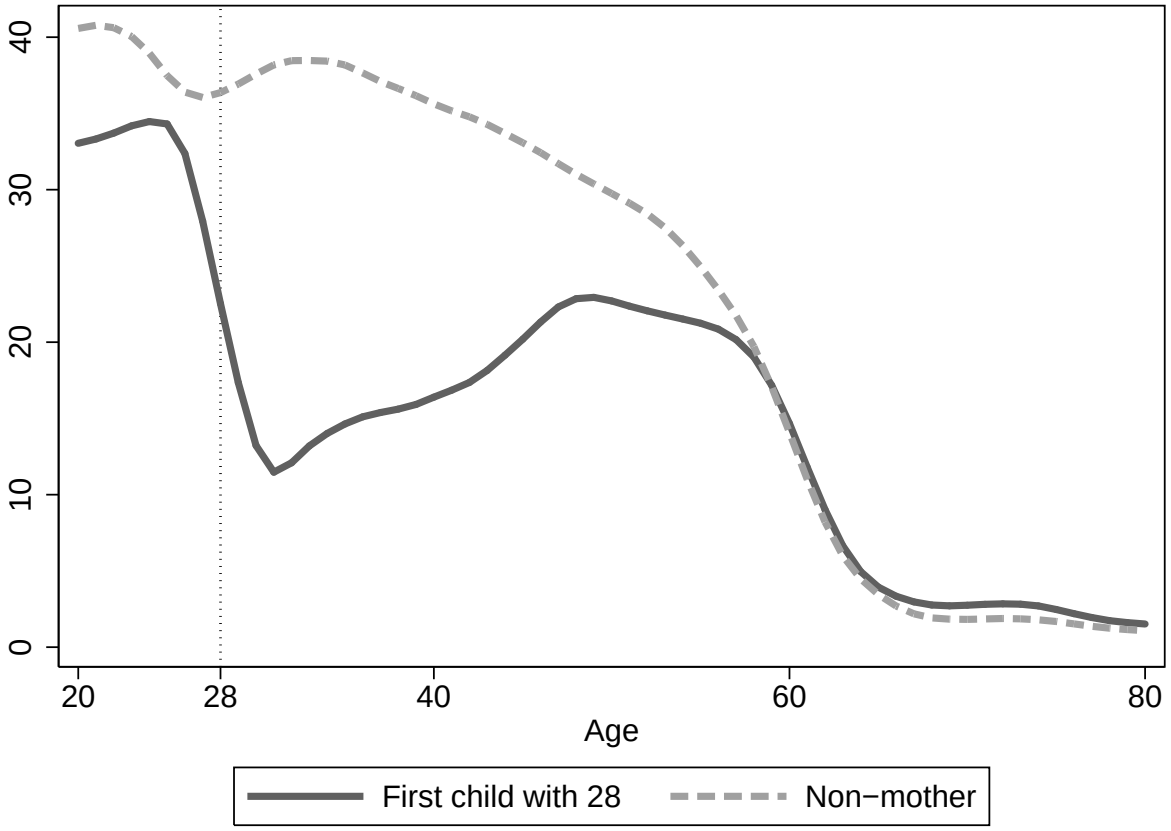
Table B.1: Regression coefficients of Figure 4.4

| | Equation (4.2) | | Equation (4.3) | | Equation (4.4) | |
|---|---|---|---|---|---|---|
| Years after first child's birth: | | | | | | |
| -1 | 0.20 | (0.07) | 0.18 | (0.10) | 0.23 | (0.07) |
| 0 | 0.56 | (0.07) | 0.52 | (0.10) | 0.56 | (0.08) |
| 1 | 0.44 | (0.07) | 0.40 | (0.11) | 0.41 | (0.08) |
| 2 | 0.04 | (0.07) | 0.11 | (0.10) | 0.16 | (0.09) |
| 3 | 0.14 | (0.08) | 0.12 | (0.11) | 0.13 | (0.09) |
| 4 | 0.05 | (0.08) | 0.03 | (0.11) | 0.03 | (0.10) |
| 5 | 0.12 | (0.09) | 0.11 | (0.11) | 0.08 | (0.10) |
| 6 | 0.09 | (0.09) | 0.07 | (0.11) | 0.06 | (0.11) |
| 7 | 0.08 | (0.10) | 0.07 | (0.11) | 0.06 | (0.12) |
| 8 | 0.08 | (0.11) | 0.08 | (0.12) | 0.02 | (0.12) |
| 9 | 0.08 | (0.11) | 0.08 | (0.12) | 0.03 | (0.12) |
| 10 | 0.12 | (0.12) | 0.17 | (0.12) | 0.05 | (0.13) |
| 11 | 0.09 | (0.13) | 0.10 | (0.13) | 0.00 | (0.14) |
| 12 | 0.14 | (0.14) | 0.13 | (0.14) | 0.04 | (0.14) |
| 13 | 0.12 | (0.15) | 0.14 | (0.14) | 0.05 | (0.14) |
| 14 | 0.13 | (0.18) | 0.15 | (0.16) | 0.04 | (0.15) |
| 15 | 0.27 | (0.19) | 0.12 | (0.17) | 0.06 | (0.15) |
| 16 | 0.05 | (0.22) | 0.10 | (0.18) | 0.03 | (0.16) |
| 17 | 0.12 | (0.23) | 0.27 | (0.20) | 0.07 | (0.16) |
| 18 | 0.27 | (0.26) | 0.19 | (0.20) | 0.06 | (0.17) |
| 19 | 0.44 | (0.31) | 0.26 | (0.22) | 0.06 | (0.17) |
| 20 | 0.25 | (0.59) | 0.36 | (0.29) | 0.05 | (0.18) |
| Number of observations | | | 37,616 | | 78,470 | |
| Number of clusters | | | 1,590 | | 9,791 | |
| F-statistic | | | 5.74 | | 14.37 | |

*Note:* The table shows the point estimates of the motherhood effect for different estimations strategies (equation (4.2): Matching; equation (4.3): Regression using past satisfaction levels and trends; equation (4.4): Fixed effects regression accounting for the anticipation effect). Cluster robust standard errors in parenthesis. The estimates are graphically presented in Figure 4.4. F-statistic for the hypothesis that all shown coefficients are equal to zero (critical value at the 1% level at 1.85).

# C Additional figures

Figure C.1: Weekly working hours of women over the life cycle




*Notes:* Data from the GSOEP waves 1984-2009 is detailed in Appendix A1. Displayed average life satisfaction paths are conditional on sets of dummies for survey years and years in panel, smoothed (Lowess) with bandwidth 0.12.

# Curriculum Vitae

## Personal Information

Date of Birth: January 14, 1985

Citizenship: Swiss

## Education

Doctoral studies in Economics, University of Zurich, Switzerland, September 2009 - April 2013

Master of Science in Economics, University of Zurich, Switzerland, April 2011

Bachelor of Science in Economics, University of Geneva, Switzerland, September 2007

## Research and teaching fields

Primary: Microeconometrics, applied econometrics.

Secondary: labor economics.